EXAMINING JUDGEMENTS:

Theory and Practice of Awarding Public Examination Grades

Michael John Cresswell

Thesis submitted for the degree of Doctor of Philosophy at the
Institute of Education
University of London
1997

ABSTRACT

This thesis reports a study of the processes by which public examination grades are awarded. Following a review of the purposes of public examinations, new theoretical analyses are given of the issues of norm and criterion-referencing, the nature of public examination standards, the problems of defining comparable standards across widely disparate assessment domains and the more technical matters of aggregating marks and examiners' judgements.

The main empirical work investigated conventional public examination grade awarding using a combination of participant observation of examiners making judgements and statistical analysis of examination outcomes. Two additional experiments are also reported; one on grade, rather than mark, aggregation methods and one on the use of strong criterion-referencing to award grades.

The main conclusions of the study are as follows:

1. Examination standards are social constructs created by special groups of judges, known as awarders, who are empowered, through the examining boards as government-regulated social institutions, to evaluate the quality of students' attainment on behalf of society as a whole.

2. As a result, examination standards can be **defined only** in terms of human evaluative judgements and **must** be **set initially** on the basis of such judgements.

3. The process by which awarders judge candidates' work is one in which direct and immediate evaluations are formed and revised as the awarder reads through the work. At the conscious level, it is not a computational process and it cannot, therefore, be mechanised by the use of high-level rule-bound procedures and explicit criteria.

4. Awarders' judgements of candidates' work are inadequate, by themselves, as a basis for **maintaining** comparable standards in successive examinations on the same syllabus. The reasons for this are related both to the social psychology of awarding meetings and to the fundamental nature of awarders' judgements.

5. The use of statistical data alongside awarders' judgements greatly improves the maintenance of standards and research should be carried out into the feasibility of using solely statistical approaches to **maintain** standards in successive examinations on the same syllabus.

6. A broadening of the range of interest groups explicitly represented among judges initially **setting** standards should also be considered.

ACKNOWLEDGEMENTS

A PhD student incurs many debts, both practical and intellectual. As recompense to my creditors I can offer only an acknowledgement of their contributions to my work, but those contributions were vital and I am extremely grateful for them all. Of course, my conclusions are my own and should not be assumed to be shared by any of the organisations or individuals mentioned below.

From start to finish, the *Associated Examining Board* has supported my work both financially and practically. I am most grateful to the Board for its generosity; to its Secretary General, John Day, for his continuing support; and to his deputy, Pat Neale, for his persistence in encouraging me when I was on the final straight.

I am pleased to acknowledge the friendly helpfulness with which John Edmundson, Secretary to the *Joint Council for the GCSE,* arranged permission for me to include the data reported in Chapter 8.

My research would have been impossible without the willing co-operation of many examiners and colleagues in subject departments within the AEB. They are too numerous to name individually, but I hope that they feel that I have done justice to their heroic endeavours to be fair to tens of thousands of examination candidates every summer. Their professionalism cannot be doubted and where I am critical it is the procedures which are my target, not the reassuringly human behaviour of those carrying them out.

The members of the AEB's *Research Advisory Committee* have offered much helpful advice and insightful comment as my work has developed. I am grateful to them all, and especially to the Committee's chairman, Jack Wrigley, for his continuing friendship and support.

My debt to my colleague Jim Houston extends far beyond his significant intellectual contributions to this project and to my thinking on assessment in general. Not only was it he who initially suggested that I should embark on a PhD but he was largely responsible for establishing the atmosphere of scholarship which marks out the AEB Research and Statistics Group. Without this, and Jim's unstinting friendship, help and encouragement, I could not have done the work reported here.

Other friends and colleagues have made innumerable vital contributions to my work on awarding, probably much more than they realise. Bob Adams, Lowena Cresswell, Martin Delap, Simon Eason, Frances Good, Hope Macdonald, Charles Newbould, Paul Newton, Martin Taylor, Murray Ward and John Wilmut have all helped me out of tight analytical spots on various occasions. I am greatly in their debt and can only hope that I have returned the favour from time to time. I am very grateful to Anita Cooper for her help with the collection and analysis of some of the statistical data and to Sarah Bawden and her colleague who so competently observed the 1993 awarding meetings on my behalf.

Harvey Goldstein has supervised my work with a masterly combination of encouragement and patience, coupled with a flair for finding the weak points in my argument which has been as helpful as it has been disconcerting. I have enjoyed our discussions and am extremely grateful for all his insight, help and friendship.

My greatest thanks go to my family: to Lowena, for persuading me that it was a practical proposition for me to do a PhD and then making it true; and to Simon and Emmeline for their forbearance when I was working and when I zipped up *Monkey Island II* to make way for Chapter Six.

Mike Cresswell
Guildford December 1996

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

WHAT IS AWARDING AND WHY STUDY IT?

"He glides upon the water's face
With ease, celerity and grace.
But if he ever stopped to think
Of how he did it, he would sink."

- *The Water Beetle* from *The Moral Alphabet*
by Ambrose Bierce

1.1 AGGREGATION AND AWARDING

In general, educational assessment involves summarising a large number of observations into a small number of indicators or, most frequently, into a single indicator. In conventional standardised testing for example, pupils might tackle 50 questions and the number of questions answered correctly by each pupil will then be counted to provide a single summary score. Similarly, in conventional public examinations the marks from different questions within a paper are added together and then the marks from the different papers are themselves added to give a total score for the examination as a whole. This process of combining observations of educational performance into a single indicator has come to be known as *aggregation*.

Once pupils' total scores have been determined, they are usually converted onto a different scale which sets their results in a more general context. For standardised tests, this may be a norm-referenced scale with known mean and standard deviation. For some criterion-referenced tests, the reported scale might simply have two values: mastery and non-mastery. For most public examinations, candidates' results are reported in terms of a scale of letter grades which is common to all the examinations within a given system (for example, all British GCE A-level examinations report in terms of the grades A, B, C, D, E and N plus an ungraded category, U). The process of converting pupils' total public examination scores into grades is known as *awarding* and is the principal subject of this thesis. In addition, some approaches to awarding involve alternative aggregation methods, so aggregation is necessarily a second important focus of the work.

1.2 PUBLIC EXAMINATION AGGREGATION AND AWARDING PROCEDURES IN OUTLINE

The aggregation and awarding practices of all the GCE boards and GCSE groups in England and Wales are similar (SRAC, 1990; SCAA, 1994; SCAA, 1995). Candidates' work is first marked by teams of standardized examiners and, since each examination typically consists of more than one component (including written papers, objective tests, teacher assessments and so on), a total score is computed for each candidate by adding together the marks from the components. Prior to addition, the marks may be multiplied by scaling factors to give the desired weight to each component within the aggregate (Cresswell, 1987a). The grading process then consists of partitioning the total score scale into contiguous regions, each of which corresponds to a particular grade. The grade awarded to each candidate is normally determined solely by the region within which that candidate's total score lies. The method by which the mark scale is partitioned uses the professional judgement of judges who are known variously as awarders or graders.

The awarders begin the job of partitioning the total mark scale by reviewing the way in which the particular examination which they are awarding has performed on the particular occasion concerned. They receive an impressionistic report from the chief examiner who has supervised all the marking and, in most boards, they review the statistics of the marks awarded. Comparisons are then made with the papers used in the previous administration of the same examination.

The awarders then scrutinize examples of candidates' work which have received different marks, judging the overall quality of each piece of work and the grade which it therefore merits. The purpose of these judgements of quality is to identify the lowest mark which is associated with scripts judged worthy of each grade. These judgements are made separately for each component and then combined (using methods discussed in detail in Chapter 4) to give an initial estimate of the lowest total mark, on the examination as a whole, which has been given to work which merits each grade.

The awarders then consider statistical evidence in the form of the distribution of grades which would follow from using the marks which they have initially identified to partition the total mark

scale into grades. This initial distribution is typically compared with the grade distribution from the previous year's examination and the expectation is that there will not be major changes from one year to the next for examinations where the number of candidates is large and the types of schools and colleges entering them are the same. Should there be large differences between the grade distributions from the two years, the awarders will reconsider some, or all, of their qualitative judgements by scrutinising further examples of candidates' work.

Having considered both the candidates' work and the statistical data, the awarders make a final judgement of the lowest mark which will be taken to merit the award of each grade for the particular examination concerned. The relative weight given to the statistical data on the one hand and to the scrutiny of candidates' work on the other varies somewhat from examining board to examining board but all of them pay some attention to both sources of information. The threshold marks for each grade are commonly called the *grade boundaries* and, as a set, define the ranges of total marks corresponding to the grades.

Once the grade boundaries have been fixed, there is a final stage of awarding called the *grade* or *borderline review* in which the results of some individual candidates are checked. These candidates are selected using a variety of criteria but generally include those candidates whose total marks fall just below one or more grade boundaries. At the start of the present study, borderline reviewing practice varied considerably both in selection criteria and in how the reviews were conducted. In some boards, the selected candidates had their work as a whole qualitatively reviewed and the grade implied by their total mark and the agreed grade boundaries was replaced by that resulting from the holistic review if the two differed. In other boards, the work of the selected candidates was re-marked to ensure that they had not just failed to be awarded a higher grade as a result of errors of marking. Combinations of these approaches were also used. However, by 1995 the examining boards' reviewing procedures all involved the re-marking approach (SCAA 1994 and 1995). The research reported here explicitly excluded borderline reviewing because it is additional and subsequent to the main focus of interest - the awarding process. A theoretical and practical study of borderline reviewing was carried out by Cresswell (1986a).

1.3 IMMEDIATE FOCUS OF THE STUDY

The present study is primarily concerned with the awarding process (prior to borderline reviewing) as it operates in British public examinations; particularly, but not exclusively, the way in which the crucial qualitative judgements are formed. From an analysis of practice and observational data, the study aims to:

1. develop a better theoretical basis for public examination awarding;
2. investigate the nature of the qualitative judgements upon which it depends;
3. investigate the information used in current awarding procedures and the way in which it is used;
4. explore possible alternative approaches to awarding public examination grades;
5. evaluate, on the basis of 1 to 4, current awarding procedures and, if appropriate,
6. recommend changes to current procedures.

Because awarding is not independent of the aggregation processes which precede it, the study also addresses aggregation methods in some detail.

The first aim may appear surprising since public examinations have operated in Britain for many decades and have, throughout most of that time, involved awarding procedures of various kinds. Nonetheless, the theoretical underpinning of awarding has never been made explicit by its practitioners. Indeed, the need for such an underpinning has not, apparently, been generally recognised. Among the very large number of books and articles published during the last 30 years on British public examinations, only one has attempted seriously to discuss the theoretical basis of awarding, although many of the rest have, perforce, made reference to the theoretical and philosophical difficulties related to setting standards which are inherent in the process. Like many other aspects of British public examinations, awarding procedures have evolved in response to changing, but essentially practical, demands made upon the public examination system. Present awarding procedures therefore have an essentially pragmatic basis, being more the product of evolution than the application of any coherent theory.

The exceptional book is that of Christie and Forrest (1981) who presented two possible theoretical frameworks and argued that the choice between them depends upon the view

taken about the function of public examinations. In this respect, although it takes issue with their frameworks, the present study builds upon Christie and Forrest's work and the functions of public examinations are considered in detail in Chapter 2.

One of the main thrusts of Christie and Forrest's book was to challenge conventional aggregation and awarding procedures by proposing that different facets of the assessed attainment domain should be explicitly considered and, perhaps, given variable weights depending upon the grade being awarded. This approach is discussed later in Chapter 4, where aggegation and its fundamental interaction with awarding is considered in depth. Empirical work on two unconventional awarding procedures which involve novel approaches to aggregation is reported in Chapters 7 and 8.

Unlike any previous work, the present study includes a detailed theoretical analysis of the fundamental nature of the judgements of quality upon which awarding depends, and the standards which these judgements define. This analysis is set out in Chapter 3. In Chapters 5 and 6 empirical work on the nature of the qualitative judgements and the way in which other data are used alongside them is reported. This work was carried out at one examining board and involved systematic observation of awarding meetings and analysis of the statistical data used in the meetings. To facilitate this study of the essential nature of awarding, the fieldwork focussed only upon GCE A-level examinations since these typically have less complex structures than GCSE examinations. The unique aspects of awarding which follow from the complexity of GCSE examinations were the subject of a three year research project by Good and Cresswell (1988a and 1988b) but the fundamental issues are the same for both GCE and GCSE examinations. Indeed, these issues arise in any large scale assessment system.

1.4 WIDER RELEVANCE OF THE STUDY

During the 1980s, there was a growing belief among policy makers that the basis for awarding public examination grades should, and could, be made more explicit. With the introduction of GCSE examinations, considerable efforts were made at the government's instigation (DES, 1982) to develop and use explicit criteria in place of conventional qualitative judgements for

awarding GCSE grades (SEC, 1985a and 1985b). This proved impossible, however, for reasons which are discussed in detail later (Chapter 3) when research on grade criteria is reviewed.

Subsequently, attempts were made, despite this earlier failure, to avoid the use of judgements of quality in the process of awarding levels to pupils taking National Curriculum assessment tasks and tests between 1991 and 1993. These attempts led to the adoption of a series of extremely complex marking and aggregation systems which, as Cresswell (1994) demonstrated, still necessarily required similar qualitative judgements to be made (albeit during test construction, rather than during the processing of pupils' marks) while at the same time compromising the reliability and validity of the assessments. By December 1993, the centrality of awarding judgements in National Curriculum assessment was being officially acknowleded in the form of a seminar organised by the *School Curriculum and Assessment Authority* for which papers on more conventional standard setting techniques and issues were invited (Cresswell, 1993a; Lundy, 1993; Massey, 1993; Morrison *et al* 1993a and 1993b; Pollitt, 1993; and Wiliam, 1993).

Much of the rationale for the work on GCSE grade criteria and National Curriculum assessments which was carried out in the 1980s and early 1990s was provided by the development of a corpus of ideas constituting what can be called *strong criterion referencing.* (Strong criterion referencing is defined, and a detailed critique of it offered, in Chapter 3.) The absence of an adequate theoretical understanding of the awarding process in general, and the nature of the judgements of quality involved in particular, undoubtedly facilitated this ill-fated development and encouraged the successive, unsuccessful attempts to implement assessment procedures based upon it.

Of course, there are major philosophical problems which beset attempts to define the notion of quality and analyse judgements of it. Pirsig (1974 and 1991) concluded that quality can be understood only in a metaphysical way and, similarly, there has been a tendency in the past for those professionally involved in public examinations almost to celebrate the obscurity of

the process of making qualitative awarding judgements. The quotation[1] from Ambrose Bierce's *Moral Alphabet* which opened this chapter neatly summarises this attitude which appeals to a mysterious capacity for making judgements of quality which is deemed to exist by virtue of the common humanity of the awarders, informed by special knowledge of examination "standards".

The present study is an attempt to begin the task of building a more analytical understanding of the function and nature of the judgements of quality which are intrinsic to the process of awarding examination grades. The centrality of such judgements in other large scale external assessment systems means that, although the study is focussed on the particular context of public examinations, its findings also have a much wider relevance (see Chapter 3).

[1] I am grateful to Gerry Forrest for first drawing the aptness of this quotation to my attention.

CHAPTER 2

THE FUNCTION OF PUBLIC EXAMINATIONS

... function is smother'd in surmise ...
- Macbeth, Act 1 Scene 3

2.1 INTRODUCTION

Many authors (for example: Ingenkamp, 1977; Christie and Forrest, 1981; Broadfoot, 1986) have discussed the possible functions of examinations in general and there is reasonably good agreement about them. Summarising this work, Cresswell (1995) lists the following various functions which examinations may be expected to serve:

> providing formative information about students concerning their progress so as to improve their future learning;
>
> motivating students;
>
> providing information for assessing the effectiveness of teaching methods, curricula, forms of organisation or schools;
>
> providing information about the performance of the education system as a whole;
>
> providing curriculum control.
>
> providing information for selection, in order to distinguish between students with different abilities and achievement so as to provide appropriate further education, training or specific employment;

In this chapter, the relevance of each of these functions to public examinations is considered and some important implications of the predominant function - providing information for selection - are addressed.

2.2 PROVIDING FORMATIVE INFORMATION

The provision of formative information does not seem at first sight to be particularly relevant to public examinations which are taken at the end of courses and are primarily summative in intent. However, British public examinations are taken in a temporal sequence and, despite their origins as school leaving examinations, are frequently taken by students who remain in full time education afterwards. Interpreting the notion of formative assessment widely, GCSE

results are thus used formatively in determining the A-level courses a student takes or in determining what he or she studies at a lower level (including GCSE re-sits themselves). A-levels play a similar role with respect to students' choices in Higher Education. However, there is also a more subtle and immediate aspect to the formative influence of public examinations, as Christie and Forrest (1981) observe. Pupils' behaviour during their courses is likely to be shaped by their knowledge of the examinations which they are to take. The practice of working past papers, in particular, means that a sort of osmosis takes place in which the examinations begin to provide information to students even before they are formally taken. The reliability of such information may not be great but public examinations nonetheless have immediate formative effects whether or not this is formally part of their purpose.

2.3 MOTIVATION

The preceding comments lead naturally to the question of motivation. The status of public examinations and their role in selection processes (discussed below) can reasonably be expected to have a motivating effect. Such an effect is widely held to exist and no doubt does exist for many students. However, there are also more subtle motivational effects, again arising from the students' knowledge of the examinations which they are to take and their own likely results. Students whose expectation is of failure in a public examination will not necessarily be motivated to try to pass, particularly if they already perceive themselves as generally of low ability. Sears (1940) showed many years ago that failure is likely to lead some students to set themselves low aspirations rather than spur them on to greater efforts. Wine (1982) analysed many different studies on test anxiety and concluded that highly test anxious students, in particular, set themselves low aspirations after poor test performances. In general, it is to be expected that the motivational effects of examinations will differ with the personality of the student concerned.

Technical features of examinations can also affect motivation during the course both positively and negatively. Gipps (1990) reports a study of student attitudes to a differentiated GCSE Mathematics examination in which candidates had to choose between different combinations

of papers which limited the GCSE grades which they might be awarded. The effect was to motivate some students and de-motivate others. As with the first function of providing formative assessment, it seems that a variety of motivational effects, some good and some bad, flow from public examinations even if positively motivating students is generally held to be one of their purposes.

2.4 PROVIDING INFORMATION FOR ASSESSING EFFECTIVENESS

This function of examinations concerns their use as measures of outcomes from differing schools, organisational strategies, teaching methods and so on. Public examination results have often been used in this way, particularly in work on school effectiveness. The statistical issues surrounding this use of examination results are complex and outside the scope of the present study. Gray *et al* (1986) reviewed them in detail and concluded that multi-level models of the type developed by Goldstein (1995) and others are required. Such models facilitate the appropriate contextualisation of schools' examination results in terms of input measures and other explanatory variables and are now almost routinely used in this context (see, for example, Nuttall, *et al*, 1989; Goldstein *et al*, 1993 or Gray *et al*, 1995). However, the statistical problems are by no means solved. Guskey and Kifer (1989) showed that the rank ordering of LEAs and school districts using such models is highly sensitive to minor changes in the variables used as input measures. Goldstein (1991) pointed out that this is to be expected because there is always a wide choice of variables which might be included in the models. This is certainly true in the case of school comparisons and Goldstein *et al* (1993) argue that while identification of extreme cases might be possible, apparently precise rank-orderings of schools from multi-level models may be seriously misleading. It is also worth noting in this connection that Nuttall *et al* (1989) reported that schools were ranked differently for different sub-groups of pupils. It clearly follows that a single ranking of schools in terms of effectiveness is of little practical use for selecting schools for individual pupils and, moreover, that schools' positions in such a ranking will depend upon the mix of pupils which they contain.

A further major issue which surrounds this use of examination results is their appropriateness as output measures. Public examinations cover only a limited range of the intended outcomes

of secondary education and, despite the high status given to these particular outcomes, therefore present only a partial picture of students' achievements. This is all the more true because not all pupils take public examinations; effective or ineffective education of these students does not, therefore, affect the examination results of a school although the proportion of such pupils in a school does affect its results. Despite the problems of interpretation, however, the examination results of schools in England and Wales are now routinely published and are widely taken, in uncontextualised form, to be indicators of effectiveness. The inadequacy of this practice does not prevent it from being seen by some as a proper use of the public examination system.

2.5 PROVIDING INFORMATION ABOUT THE EDUCATION SYSTEM AS A WHOLE

Christie and Forrest (1981) proposed this function which is clearly related to the effectiveness function discussed in the previous section. Certainly, it shares the same problem of the partial nature of the information provided by public examinations. Christie and Forrest argue that public examination results could be used to monitor the standards of the education system over time if awarding involved the use of explicit criteria. It will become apparent in later chapters, however, that this view is näive about the nature of such criteria. It also ignores the logical problem of measuring quantitative change in attainments in a qualitatively developing curriculum (see Goldstein, 1983 and Nuttall, 1986). Nonetheless, examination results are often taken by policy makers and others to be indicators of national standards; witness the inclusion of examination results in the Department of Education and Science's annual statistics publications and the press coverage given to each year's results when they are issued.

2.6 CURRICULUM CONTROL

Several authors, for example Broadfoot (1986) and Tattersall (1994), have argued that the public examination system in England has served as an instrument of curriculum control. In a narrow sense, this is true but it is also true that examinations have responded to curriculum change and a system of complex interactions is a more realistic model of the historical

relationship between the two. Nonetheless, those who have devised public examination syllabuses have effectively defined the courses of study to be followed by pupils in most secondary schools in England and Wales although there has generally been considerable choice from among competing syllabuses. More fundamentally, the very existence of public examinations and their administrative and technical requirements have influenced school practice profoundly (Mortimore and Mortimore, 1984; Mathews, 1985). Examinations have provided, by virtue of some of the functions discussed above, clear goals at which schools have aimed, sometimes almost to the exclusion of all others (DES, 1979). The public examination system's structure of separate academic subjects has exerted considerable influence over the organisation of schooling. Inclusion in the examining system has conferred status upon curriculum areas which achieve it (Apple, 1978) and this mechanism has had considerable effects upon pedagogical practice outside the mainstream of academic subjects.

In the light of all this, it is useful to try to identify the agencies which have been exercising such control. At 18+ and to a lesser extent 16+, GCE examinations provided a major channel through which the universities influenced the school curriculum. In CSE examinations, on the other hand, it was usually groups of involved teachers who determined the content of the syllabuses and, to a lesser extent, the assessment techniques used. Historically recent developments such as Mode 3 examinations gave individual teachers substantial freedom to devise their own courses, albeit within the overall public examination structure. In fact, as was so often the case in the British education system in the past, it is difficult to identify a single controlling agency. As Mathews (1985) has argued, the power and influence exercised through the examination system was, historically speaking, dispersed between many different agencies.

However, the 1988 Education Reform Act caused a considerable shift in the balance of power between these various agencies. In particular, the role of central government was made pre-eminent. The legally defined role of the School Examinations and Assessment Council and its successor body the School Curriculum and Assessment Authority in approving syllabuses and qualifications has placed the control mechanism of public examinations at 18+ more firmly in centralised hands than ever before. At 16+, the curriculum is now legally specified and public

examinations have become the method by which candidates are assessed at the end of a centrally devised programme of study rather than the principal external device for defining such programmes. It is worth noting, however, that the requirements for assessment which are a part of the National Curriculum restrict the options of those specifying programmes of study. In this sense, technical and administrative features of large-scale examinations and other assessments continue to influence the school curriculum. The operation of this effect was evident in the review of the National Curriculum carried out in 1993 (Dearing, 1994).

2.7 PROVIDING INFORMATION FOR SELECTION

The role of public examinations in providing information for selection has motivated much of the concern that has been expressed about the quality of the information which they provide. Ever since Edgeworth's pioneering study (1890) there has been a considerable amount of research into the technical features of examinations. However, the quality of the selection decisions to which the examination results contribute is of more practical significance. This depends not only upon the reliability and validity of the examination results but also upon the quality of any other information which is also used and, of course, upon other features of the selection processes themselves. Goacher (1984) reports that:

> "A range of other selection criteria appeared to be applied by both industrial and educational users in a confident if rather haphazard way. While rejecting teacher assessments carefully built up over time (and often proffered with caution as to their subjectivity), many users claimed that they could elicit details of personality and behaviour in a brief interview. In educational settings there was little evidence of preparation for or co-ordination of this type of activity and while large companies may well train their selectors, it is doubtful if such training is undertaken by those working for smaller firms or educational institutions."

Given this, it is likely that, with their known levels of reliability (Willmott and Nuttall, 1975; Murphy, 1978 and 1982a; Newton, 1996) and predictive power (Miles, 1979; Houston, 1987; Kellaghan, 1995), public examination results are among the better sources of information used by selectors.

Intimately connected with the role of public examinations as part of selection processes is the effect which they have of legitimising those processes. As Broadfoot (1979) says:

> "On the one hand, it is possible to see the institution of various kinds of educational assessment as crucial steps in the fight against nepotism and inefficiency and in opening up social mobility to a quite unprecedented extent. On the other hand, it is important to recognise the role of assessment in limiting such mobility and even more crucially, in legitimating what is still an education system strongly biased in favour of traditional privilege."

Certainly, there is no doubt that some social groups are disproportionately represented among those who are successful in public examinations (see Mathews, 1985 for example) and to the extent that examinations are, nonetheless, believed to be essentially meritocratic, they will have a legitimising effect. This issue is closely connected to that of examination bias which is briefly discussed later in this chapter.

2.8 THE PRE-EMINENCE OF THE SELECTIVE FUNCTION AND THE IMPLICATIONS OF THIS FOR THE AWARDING PROCESS

There is little doubt that the selective function of public examinations is the predominant one (Cresswell, 1995). It is identified as the "principal historical function of school examinations" by Orr and Nuttall (1983); as the "historical function of the GCE examination system" by Christie and Forrest (1981); as the function which "outweighs all others" by Ingenkamp (1977) and as "most important" by Broadfoot (1986). Broadfoot argues that the development of public examinations in the late nineteenth and early twentieth centuries reflected a growing belief in meritocracy as a social organising principle, coupled with the notion that it was appropriate for schools to provide for the selection which such an ideology implies. Both of these beliefs are still evidently widespread in the late twentieth century so it is perhaps not surprising that examinations continue to have a major role in most educational systems (Kellaghan, 1995).

While the nature of examinations has changed considerably over the years, recent evidence supports the idea that the selective function remains the primary one. Goacher (1984), reporting a two year study of the uses made of public examination results, observed (p 124-5):

> "The study indicated that the uses made of examination results by industry and within education were not as distinctive as might have been expected. Some subjects inevitably had an almost vocational value in educational settings that they lacked in industry; otherwise, examinations were viewed and used in a remarkably similar way ...

> Examinations were principally used to control numbers. They were used almost universally as a mechanism to screen out excess applicants, with the cut-off level being a function, among other factors, of the ratio of applicants to places. The association of entry level with course status was also not without influence ...
>
> Having been used to screen out excess numbers, examination results were then usually used to rank candidates to facilitate selection."

As noted in Sections 2.4 and 2.5, since the time of Goacher's study there has been a considerable increase in the use, however questionnable, of British public examination results as indicators of the effectiveness of individual schools and the educational system as a whole. Nonetheless, Bishop *et al* (1996) confirm that the selective function remains the one which is seen, by parents, teachers, employers and candidates themselves, as the *raison d'etre* of public examinations.

If selection is the principal purpose to which public examination results are put, then the issue of fairness is obviously crucial. Given that a meritocratic philosophy underpins the use of examination results for selective purposes, it is appropriate to judge fairness in this light. There are a number of conditions which an examination must meet to be fair in meritocratic terms. The examination must be unbiased, comparable, reliable and valid. These conditions are briefly considered in turn below and the implications of them for the awarding process are discussed.

2.8.1 Bias

The issue of bias in assessment is a complex one and the reader is referred to Gipps and Murphy (1994) for a full discussion. From the meritocratic point of view, it is undesirable for examination grades to be biased against any particular social group since it is precisely the social biases evident in other selection procedures, such as patronage, which meritocracy is intended to eliminate. However, the problems of bias extend beyond the examinations themselves. It is well established that social variables, race and gender are all related to school achievement and, given the social nature of schooling, members of different social groups have different educational experiences which are likely to produce different measured attainments in examinations (Mathews, 1985; Gipps and Murphy, 1994).

It is sometimes argued (see, for example, Dore, 1976) that, to be fair, selections should reflect some underlying ability or aptitude, rather than current attainment which is more open to the effects of social inequalities. Nuttall and Willmott (1972) argued that, from this perspective, a single general intelligence test could be used in place of public examinations for selection purposes. However, Wood (1986 and 1991) clearly sets out the theoretical problems associated with aptitude testing and shows that the available evidence suggests that, in practice, aptitude assessments are no more free of the effects of environment and school-mediated social inequalities than assessments of attainment. Moreover, their use for high-stakes assessments can have undesirable backwash effects upon subject teaching and learning.

To take a concrete example of apparent bias relating to attainment, socially disadvantaged pupils, as a group, are proportionally under-represented among those with good public examination grades (see, for example, Blackstone and Mortimore, 1982). The question about bias is: should they be under-represented in this way? Mathews (1985) argues strongly that they should, because examination grades should reflect current attainment and socially disadvantaged pupils have, as a group, relatively low attainment. Similarly, Gipps and Murphy (1994) argue that fair assessments must meet a principle of equity but that this does not require equality of outcomes. This position is consistent with the meritocratic perspective.

Of course, there remains the question of what attainments examinations should report and how these should be assessed. The position which has traditionally been taken by the public examining boards is that the attainments measured by public examinations, and the assessment techniques used, should be determined without explicit reference to their effect in terms of the performance of different social groups. Rather, what is being assessed should be determined on pedagogical grounds and how it is assessed is a technical question involving finding the best techniques for the purpose. Examining board procedures then involve simply making every effort to remove superficial signs of bias such as any preponderance of, say, masculine contexts for problems. Once this has been done, any remaining differences in the outcomes for different subgroups of candidates for public examinations are seen as a

reflection of the way in which subjects are defined. Clearly, this position begs many historical questions concerning why subjects are defined as they are and why particular assessment techniques are regarded as best. For example, it is clear that changes in the extent to which coursework contributes to an examination will change the measured difference between boys' and girls' attainments (Stobart *et al*, 1992; Cresswell, 1993b). Similarly, there is evidence that multiple choice tests produce different gender differences from other written examination components (Murphy, 1982b). Using effects of this type, it would be quite possible to organize examinations so as to eliminate average differences between boys' and girls' measured attainments. Whether this would be desirable, however, is a social and ethical question, rather than a technical one (Goldstein, 1996).

As regards gender effects at least, the British legal system takes a similar view to the examining boards. The old practice of using separate standardisation tables for boys and girls in selection tests is now illegal. As Goldstein (1986a and 1993) has pointed out, to use particular assessment techniques or to assess particular attainments in order to engineer equal outcomes for boys and girls would be no different, in effect, or spirit, from using different standardisation tables for boys and girls. It would seem, therefore, that to take social factors into account during awarding (for example, to locate the grades on the mark scale in different places for different social groups) would not be a generally acceptable practice. Certainly, it is never done and therefore, as far as the present study of awarding *per se* is concerned, the problems of examination bias need not be considered further.

2.8.2 Comparability

The emphasis which has traditionally been placed upon studying the comparability of different examinations reflects the centrality of the selective function of public examination results. Selectors often treat grades from different examinations as interchangeable and this is fair from a meritocratic point of view only if the examinations concerned are comparable in the sense that the same grade represents equal merit regardless of the examination from which it comes. For example, an employer might specify a requirement of 5 Grade C GCSEs including English and Mathematics. This clearly assumes that a Grade C in English from any

syllabus, taken on any occasion, has equal merit to a Grade C from any other syllabus on any other occasion, and the same for Mathematics. It also assumes that Grade Cs in all other subjects are of equal merit.

The meaning of comparability between public examination grades is therefore at the heart of the awarding process and it is explored in depth in later chapters. Here, it will simply be noted that there are good reasons for believing that it is impossible to establish whether or not examinations are comparable in the way in which the term is normally understood (Goldstein, 1986b; Cresswell, 1996; Goldstein and Cresswell, 1996). In fact, several ways of defining equal merit have been implicitly used by those studying examination comparability and these are discussed in Chapter 3 where a new definition of comparability is proposed which does not suffer from the theoretical contradictions and circularities inherent in previous definitions.

2.8.3 Reliability

A necessary feature of fair meritocratic selection systems is that the selections should be replicable: ideally, if the same individuals are considered on a different occasion, the identical subset of them should be selected; in practical terms, this requirement should be met on average over different replications. Selection systems which do not have this characteristic, such as random lotteries, cannot be fair in meritocratic terms. It follows that the same individuals examined on different occasions must receive the same results on average if subsequent selections are to be fair. Examination results therefore need to be as reliable as possible if the selection decisions based upon them are to be as fair as possible.

Public examination grades are usually awarded by first marking candidates' work and then partitioning the mark scale into bands which each correspond to a particular grade. The reliability of the grades therefore has two principal determinants: the reliability of the marks and the reliability of the awarding process by which the mark scale is partitioned. The reliability of marking in general has been extensively studied. Wilmut (1986) compiled an extensive bibliography of over 240 studies or reviews relating to it. Murphy (1978 and 1982b)

and Newton (1996) both reported generally high levels of marking reliability in modern public examinations.

Far less attention has been paid to the reliability of the public examination awarding process. It appears that only one study has attempted to assess it directly. In that study, Good and Cresswell (1988a and 1988b) concluded that the level of reliability achieved was acceptable because it was equivalent, in terms of the effect upon candidates' grades, to marking reliabilities of 0.96 or better. Indirect evidence about the reliability of the awarding process comes from some of the many studies of grade comparability between different examinations which have been carried out. In general, the results of these studies, which have been reviewed by Bardell *et al* (1978) and by Forrest and Shoesmith (1985), have been consistent with the hypothesis that examination grade awarding is reasonably reliable. The reliability of the awarding process is considered further in Chapter 3.

2.8.4 Validity

The use of public examination grades for selective purposes pre-supposes that they are positively related to success in the job or course concerned. There is little published British evidence which directly concerns this relationship in the vocational field but the matter has been studied in the context of selection within the world of education. Houston (1987) reviewed the large amount of work which has been done on the relationship between GCE A-level grades and performance in higher education. In particular, he cited a major study by Choppin and Orr (1976) as demonstrating that, of the available measures which might be used for this selective purpose, GCE A-level grades were the best single predictor of future success. No better single predictor had been identified in any of the studies which Houston reviewed. Forrest (1995) reported the same conclusion and similar results were also found by Miles (1979) in a study of the ability of GCE O-level grades to predict later results at A-level. At a recent international conference on selection for Higher Education (Kellaghan, 1995), predictive validity coefficients in this context were reported from several different countries and few exceeded 0.5 with none exceeding 0.6. The clear implication of the available evidence in this area is that only weak predictions of future performance are possible. This is hardly

surprising and although the validity of public examination results for selective purposes may be **relatively** high, in absolute terms examination results do not give rise to accurate predictions of future performance.

However, there is a sense in which this is to miss at least part of the point. The meritocratic philosophy which underpins the use of examination results for selective purposes contains a strand which holds that future rewards for an individual should, for moral reasons, reflect that individual's current achievements - that success must be earned (Young, 1961). In response to this it might be argued that examinations measure attainment rather than achievement and that for some individuals a relatively low attainment may represent considerable progress after the investment of great effort which, for similar moral reasons, should be rewarded in some way. Indeed, it may be that rate of progress is as good a predictor of future success as level of attainment. However, the meritocratic view is that high **attaining** individuals are also more deserving of selection than others on the utilitarian grounds that they are better equipped, by virtue of their current attainment, to be successful if selected. Note that it is possible to hold this view despite the low values usually reported for predictive validity coefficients since, for practical reasons, these generally relate to differences of performance between **selected** individuals and are therefore attenuated to some unknown extent. Finally, selectors, particularly in education, usually argue that the level of attainment reached by a new student needs to exceed some pre-requisite (if loosely defined) level if they are adequately to deal with the demands of their new course. This, too, points to a general view that **current attainment** should be the concern of assessments intended to assist selection.

From this perspective, the validity of public examination grades as measures of current attainment is important to their role in social selection. There is, however, scant empirical evidence of the validity of public examinations in this sense. Because public examination candidates rarely take more than one examination in a given subject on one examining occasion, there are no data sets of any size which might give correlational measures of concurrent validity. The strong claim to validity made by public examinations rests principally on their content. As Messick (1987) indicates: "Content validity is based on professional judgements about the relevance of the test content to the content of a particular behavioural

domain of interest and about the representativeness with which the item or task content covers that domain". Examining boards place great procedural emphasis upon establishing relevance and representativeness in this sense but the only external evidence of their success in this regard is the continuing acceptance by most teachers that the content of most public examinations is reasonably appropriate to the courses which they teach. However, such acceptance does continue and represents a substantial body of professional judgements suggesting that most public examinations exhibit an acceptable degree of content validity.

As far as the awarding process is concerned, the need for content validity has only indirect implications. Content validity is a prior matter affected by the design of papers and assessment schemes. It is argued in Chapter 3 that these matters have very important implications for the definition of the standards which underpin grade awarding. However, the awarding process *per se* does not influence content validity. Predictive validity is potentially a different matter, however, because the use of more grades would produce higher measured predictive validity (Shaw *et al*, 1987). Nonetheless, Cresswell (1986b) has pointed out that the question of how many grades should be used to report public examination results is not purely, or even primarily, a technical one. It is therefore unlikely to be much affected by validity considerations and is effectively pre-determined as far as the awarding process in practice is concerned. The work of Shaw *et al* (1987) implies that the particular locations of a given number of grades upon the mark scale will have little effect upon the information loss incurred by awarding and therefore little effect upon the predictive validity of the grades.

2.8.5 Level of generality

The level of generality of the information provided by public examinations is also an important issue with respect to selection. It is clear from Goacher's (1984) study, and from other evidence (for example, Harrison, 1983), that selectors do not wish to make complex selection decisions involving simultaneous consideration of several different aspects of attainment within a subject, instead they prefer a single summarising measure - a subject grade. It is, of course, possible to argue that this is because selectors are naïve about the task which they have to perform and that only a single summarising measure is usually provided. However,

there is some theoretical justification for selectors' preferences in this matter. Making decisions on the basis of several attributes at once is notoriously difficult and much of social judgement theory is concerned with attempts to explain the practical behaviour of judges in terms of the relative values which they attach to the various relevant attributes of the objects being judged (see, for example, Hammond *et al*, 1975). From this perspective, the use by selectors of a single summarising measure of educational attainment in each subject - an examination grade - is rational given that they are content with the relationships between the different aspects of attainment which are implied by the aggregation procedures used to derive it. Even then, the substantial problem still remains for selectors of appropriately combining examination grades in several subjects with other information about prospective employees or students. The implications for the examining process of the level of generality of examination grades relate directly to methods of aggregation and indirectly to awarding. These implications are discussed in more detail in Chapter 4.

2.9 CONCLUDING REMARKS CONCERNING THE SELECTIVE FUNCTION OF PUBLIC EXAMINATIONS

It has been argued in this chapter that the principal function of public examinations is the provision of information to inform selection of candidates in education or employment. The rationale underlying this function is a meritocratic one. From the point of view of the awarding process with which this study is concerned, this perspective goes a long way towards explaining the traditional emphasis upon investigations into the comparability of the grades from different examinations. The primacy of the selective function has a profound effect upon the way in which the awarding process is conceived and underpins current grade standards and awarding practice in a fundamental way. In Chapter 3 some of the theoretical implications of the selective function are explored in depth; to conclude this chapter, the two major practical implications are described.

2.9.1 The need for transparent procedures

The first practical implication of the primacy of the selective function is the need for examining procedures to be accepted as fair by candidates, teachers, parents and others with an interest

in the examination results. At one time, such acceptance was forthcoming apparently without serious question. Prior to the 1980s, the examining boards were, in general, trusted as expert bodies whose judgements should only be challenged in rare individual cases. Society's attitudes have changed however, and there is an increasing reluctance to take on trust the work of expert institutions generally. Inevitably, this change of attitude applies to examining boards and to be acceptable to society at large examining systems need to be explained to those affected by them. Without dissenting in any way from the notion that such explanations should be forthcoming, it is worth observing that the requirement to explain places a limit upon the degree to which procedures can be elaborated to deal with the complexity of the assessment task. As a result, the technically best solution to an examining problem may not always be the solution chosen. For example, one of the reasons why examining boards do not scale examination component scores *post hoc* to ensure that their weights within the examination total score are exactly as intended (Cresswell, 1987a) is concern about the transparency of the procedures involved (Adams and Wilmut, 1981). As already mentioned, the use of a scale of only about 6 distinct grades to report public examination results has a number of technical drawbacks but considerable advantages in terms of perceived reliability (Cresswell, 1986b). In awarding too, a balance needs to be struck between technical considerations and the acceptability of the procedures to a lay audience. This point recurs throughout the following chapters.

2.9.2 The practical imperative

The second practical implication of the social function of public examinations is that they **must** always produce a grade for every candidate (except in rare cases of disqualification or absence) by a certain date. In awarding, this means that apparent conflicts of evidence about the grades which should be awarded **must** be resolved by a certain date, even if no other evidence can be obtained. The consequence is sometimes to force a choice between conflicting evidence when there is no direct information available to form the basis for such a choice. This feature of practical awarding procedures is particularly relevant to the discussion in Chapters 5, 6 and 9.

CHAPTER 3

A THEORETICAL PERSPECTIVE ON AWARDING AND EXAMINATION STANDARDS

"Myself when young did eagerly frequent
Doctor and Saint, and heard great Argument"

The Rubaiyat of Omar Khayyam
Edward Fitzgerald

3.1 A MODEL OF THE ASSESSMENT PROCESS

In this section, a novel model of the process of assessing pupils' work is presented. The purpose is to provide a theoretical context within which better to understand the role of awarding in public examinations. The model makes use of the notion that *description, interpretation* and *evaluation* are distinct aspects of any critical appreciation of an object. In this respect, the model draws upon the philosophy of aesthetics (see, for example, Aldrich, 1963). Although the focus of interest in educational assessment is the work of pupils, rather than a natural or man-made object which evokes an aesthetic response, there are strong parallels between the two fields.

One of the major differences between the two fields, however, is that the work of pupils which is formally assessed is generally produced in response to a particular set of assessment tasks. This is not always true of the objects with which aesthetics is concerned. As will become clear later, the fact that pupils' responses to particular tasks are the objects being assessed, is of central importance in the present study. However, the study is not primarily concerned with the detailed processes of setting such tasks and, throughout, these processes are assumed to have occurred and to have produced tasks which are content valid (in the terms set out in Chapter 2).

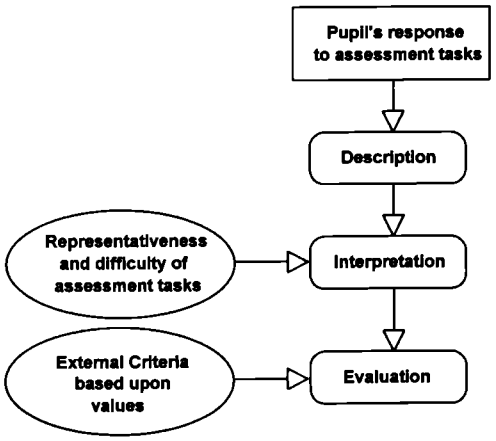
The proposed model of the process of assessing pupils' work posits three distinct types of report which assessment systems can produce, depending upon whether the information

which is reported is descriptive, interpretative or evaluative. These terms will be defined as follows:

> A *descriptive* report is a statement about a pupil which describes the attainment demonstrated by that pupil on a particular set of assessment tasks.
>
> An *interpretative* report is a statement about a pupil which interprets descriptive information in terms of a larger universe of knowledge, skills and understanding than the particular tasks to which the descriptive information relates.
>
> An *evaluative* report is a statement about a pupil which attaches a particular value to descriptive or interpretative information about them. (The more usual philosophical term for evaluative in this sort of context is *normative* but this has been avoided here to prevent confusion with norm-referencing.)

The proposed basic model of the process by which useful assessments are made is shown in Figure 3.1.

Figure 3.1
The basic description-interpretation-evaluation (DIE) model



The model has three stages corresponding to the three types of information which can be reported. First, the attainment demonstrated by the pupil on a particular set of assessment tasks is described. Second, the addition of information about those tasks enables the description to be interpreted in terms of the wider universe of knowledge, skills and understanding which the tasks are intended to represent. Finally, the interpretation is evaluated using external criteria based upon some set of values which accord merit to

attainments of a particular type or degree within the assessment universe. Note that interpretation and evaluation of the original description both require reference to information which is external to that provided by the pupil's completion of the assessment tasks.

It is important to acknowledge immediately that each stage of the assessment process is affected by the values of its designers and operators. For example, definition of the assessment universe imposes particular values by identifying the attainments which are of interest and their relative importance; selection, from that universe, of the knowledge skills, and understanding assessed in a particular instrument superimposes a second set of values; judgement of the merit of pupils' responses brings to bear a further set of values. However, statements such as *the ability to recall historical fact will be assessed* and *each fact tested is of equal importance* are logically distinct from the statement *an **acceptable** performance is the recall of 75% of the historical facts tested.* In the terms of this study, what distinguishes evaluative from descriptive and interpretative reporting is the explicit application of values of this last sort which assign merit to pupils' assessed atainments.

It must also be acknowledged that the distinctions between description, interpretation and evaluation are blurred by the hierarchical nature of practical assessment procedures in which items are grouped into questions, questions are grouped into papers, papers into examinations, examinations into certificates and so on. Moving from one hierarchical level to the next involves aggregating assessments together (aggregation is discussed in detail in Chapter 4) and therefore involves evaluative trade-offs within the examination. A conventional marking scheme in Mathematics, for example, requires markers to identify specified acceptable answers to each question and specifies the equivalence, in terms of marks, between them. Conventional aggregation processes of addition of the marks then treat the marks from different questions as interchangeable. The process of drawing up the marking scheme is thus evaluative in that it involves attaching particular relative values to different responses.

In some academic subjects (for example in English) not only is drawing up the marking scheme an evaluative process but so is its application. Markers are required not only to

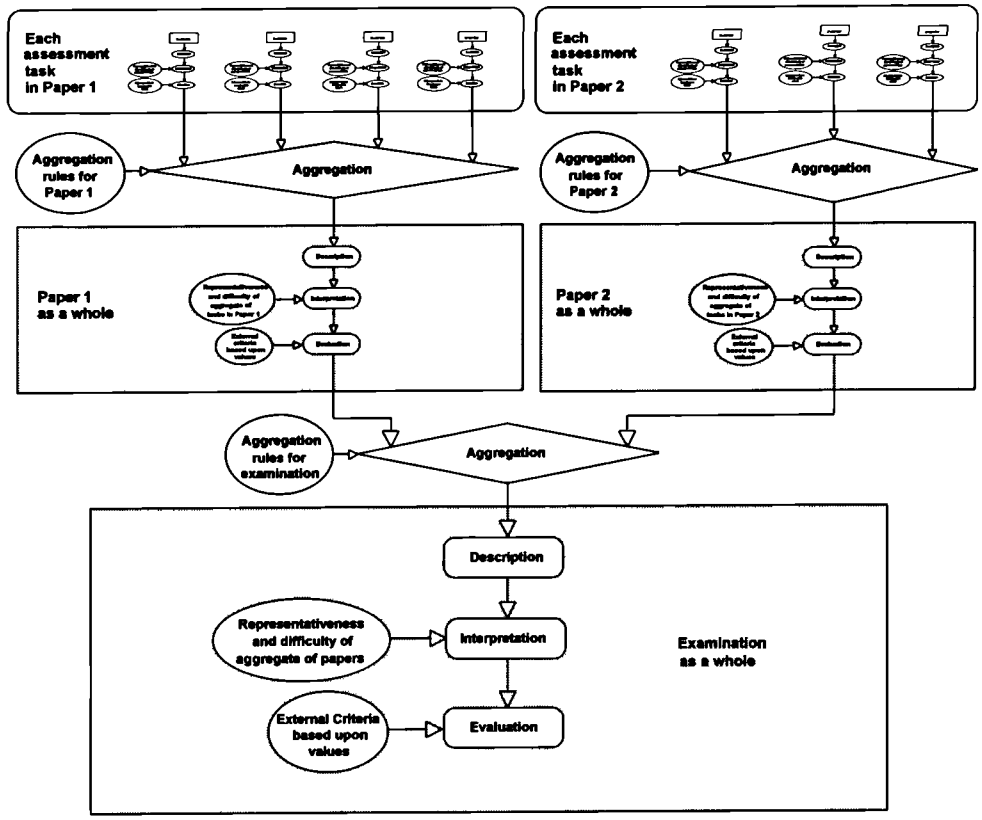
identify the occurrence of any credit-worthy features of pupils' work which are described in the marking scheme (as in Mathematics) but also to distinguish different levels of competence for each of those features. It is worth noting that this additional evaluative aspect of the marking process in English is suggested by Newton (1996) as one of the reasons why marking is less reliable in English than in mathematics.

Evaluation of some sort, then, is always involved in the descriptive stage. The distinction between this evaluation and that which is of primary interest in the present study is one of hierarchical level. Within what is the descriptive stage (ie the marking process) at the principal level of interest of this study, it is possible to distinguish the three processes of description, interpretation and evaluation. Similarly, within higher levels of the assessment hierarchy, the same three processes recur. (For example, many university entrance procedures which use A-level grades take each candidate's grade profile as a description, interpret it in terms of a points score on the subjects offered by the candidate which are relevant to the subject to be read and then evaluate this total points score against some criterion of acceptability.) The recursive nature of the hierarchy of assessment processes in a conventional public examination is represented diagrammatically in Figure 3.2.

The analysis in this chapter treats the process of determining each candidate's total examination score as being descriptive and this defines a particular level of interest in the assessment process hierarchy. However, this hierarchical level is simply the one at which the theoretical model is particularly helpful to an understanding of awarding in public examinations. It has no theoretical uniqueness.

Within the hierarchical level of primary interest, each stage in the assessment process produces a report which could be communicated to a third party - the user of the assessment results. A descriptive report, then, is one which leaves both interpretation and evaluation to the user. An interpretative report is one which leaves evaluation to the user. An evaluative report is one in which an evaluation of the measured attainment is given to the user.

Figure 3.2
The hierarchically recursive arrangement of assessment processes in a public examination



Most of the purposes of assessment set out in Chapter 2 involve the evaluative use of the assessments. For example, an assessment of a pupil carried out for formative purposes must be evaluated against that pupil's previous attainment and/or with the objectives of the course he or she is following to obtain formative information. Assessments made as part of a study of a new curriculum project will be evaluated by comparison with some control group or, much more frequently, with the aspirations and intentions of the developers of the new project. Similarly, monitoring the functioning of an entire educational system involves attempts to evaluate the current attainments of its pupils either by comparison with those of previous cohorts or by comparison with what are held to be desirable outcomes of the system. Selecting pupils involves evaluating each pupil's attainment by comparison with that of his or her peers. Thus, by itself, the descriptive score of a pupil on a particular set of assessment tasks enables few of the purposes of assessment to be met. In general, this score must be interpreted and then evaluated. It follows that the type of report which is required from a

particular assessment instrument generally depends upon the extent to which the user of the results can supply the additional information required to evaluate them. This is a key point.

In the following section, a number of different approaches to, and purposes of, assessment are considered in terms of the proposed model which is referred to hereafter as the DIE (Description, Interpretation, Evaluation) model. This analysis clarifies the relationships between public examining procedures, norm-referencing and criterion-referencing and provides the basis for the theory of awarding which is developed later in this chapter.

3.2 REFERENCING SYSTEMS

3.2.1 Conventional Criterion-referencing

Before setting conventional criterion-referencing into the context provided by the DIE model, it is necessary to review the development of the notion of criterion-referencing. The Oxford English dictionary gives both of the following definitions of a *criterion*:

1 A canon or standard by which anything is judged or estimated;

2 A characteristic attached to a thing by which it can be judged or estimated.

Originally, the term *criterion-referenced* referred to the second of these senses, as is clear from Glaser's earliest writings on the subject (1963). Glass (1978) confirms this, quoting a personal communication from Glaser in 1976 as saying that criterion-referenced tests were envisioned as being "closely articulated to the relevant behaviours which traditional psychometrics embodied in the criterion scale but seldom in the test itself". Glaser (1963) described criterion-referenced tests as measuring the "degree of competence attained by a particular pupil" on "a continuum of attainment". The emphasis was clearly on the refinement of the test instrument so that the interpretation of scores in terms of a wider universe was facilitated. Osburn's (1968) notion of *Universe defined* tests and Hively's (1970) definition of *Domain referenced* tests were, essentially, alternative formulations of the same idea.

Glass (1978) has analysed the development of the notion of criterion-referencing during the 1960s. He points out that almost from the beginning some authors interpreted the term

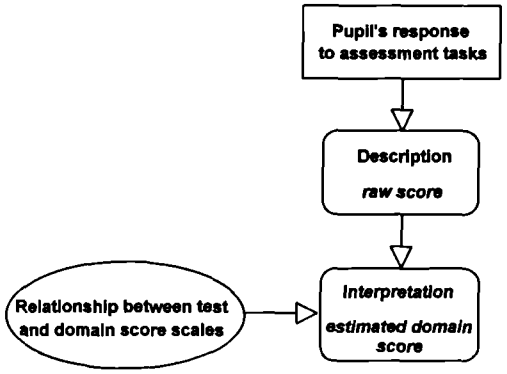
criterion in the first of the above senses as well as in the second. The idea that criterion-referenced tests should be essentially concerned with establishing mastery was, in fact, a substantial shift from the original conception. Nor did this shift have uniformly positive consequences. For example, the subsequent unsatisfactory attempts to redefine mastery using continuum models in which "mastery is seen as a continuously distributed set of abilities, and an individual's test performance places him at some point on this continuum" (Pilliner, 1979) were necessary only because Glaser's original emphasis on the continuous nature of achievement had become obscured.

It is clear, however, that despite the confusion introduced by the term criterion, both the original conception of criterion-referencing and the later one have in common the need to closely specify the content of the test. It is also clear that both conceptions assume a conventional numerical scoring process at the level of individual questions. Indeed, these two aspects are intimately connected: closely defining the domain to be assessed is intended to facilitate the selection of representative assessment tasks so that, in terms of the DIE model, the interpretation of the pupil's descriptive total score becomes immediate. In principle, a score of x% on the test will imply that the pupil can be expected to answer correctly approximately x% of all the questions which could possibly be set on the entire domain. Clearly, interpretations of this sort will be most useful, pedagogically speaking, when the domain is well-defined as it should be in conventional criterion-referenced assessment. In the terms of the DIE model, conventional criterion-referencing represents an attempt to provide interpretative assessments which are directly derived from underlying descriptive ones by clarifying, and thus making accessible external information about, the relationship between the assessment tasks and the wider universe of knowledge, skills and understanding.

In practice, the particular characteristics (such as difficulty) of the chosen assessment tasks may make it necessary to transform the raw descriptive score from a particular test so as to put it onto a common scale measuring competence on the assessment domain as a whole. A number of empirical equating techniques for doing this exist (see Holland and Rubin, 1982). In general, therefore, the interpretative report from a criterion-referenced assessment will be

an estimate of the pupil's domain score. Conventional criterion-referencing is illustrated in Figure 3.3.

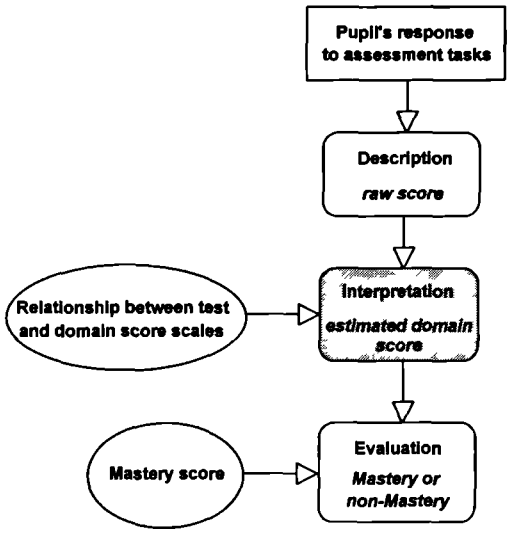
Figure 3.3
Conventional criterion-referencing in terms of the DIE model



The absence of an evaluative report from the asssessment model in Figure 3.3 is significant. It is consistent with the original notion of criterion-referenced tests as measures of continuous competence variables. Moreover, since a major intention of conventional criterion-referencing is to provide formative information, it is reasonable to argue that teachers can evaluate the interpretative report in a variety of ways to suit their particular formative concerns. For example, to measure progress by comparing the result with one from a previous assessment or to identify those pupils in greatest need of additional teaching by comparing different pupils' results with the objectives of the course.

Only when the assessment system, rather than its user, attempts to provide a view about the worth of a pupil's attainment, will an evaluative report be generated. In particular, the determination of a cut-off score denoting mastery of the domain involves evaluation of particular domain scores and produces an evaluative report as shown in Figure 3.4. Figure 3.4 also illustrates a common feature of assessment systems which produce evaluative reports. Although an interpretative stage is required if the evaluation is to have general utility beyond the particular assessment tasks used, explicit interpretative reports are not, in general, produced alongside evaluative ones. This is indicated by the shading in Figure 3.4.

Figure 3.4
Conventional Criterion-referenced mastery testing in terms of the DIE model



3.2.2 Strong Criterion-referencing

From an international perspective, criterion-referenced assessment is conceived in the ways outlined above. However, recent developments in the United Kingdom, particularly in connection with the introduction of the National Curriculum in England and Wales, took very seriously the alternative meaning of a *criterion* as a *standard*. A view became current that numerical marking procedures were incompatible with criterion-referencing. Rather, attempts were made to formulate brief verbal statements to act as standards in terms of particular competencies such as *Solve whole number problems involving addition and subtraction* (DES, 1991). In this approach, it was assumed that the identification of individual pupils' attainments with such *statements of attainment* was unproblematic; that observation of pupils solving tasks and scrutiny of their solutions was all that was required to decide for each individual pupil whether, for example, he or she **could** "solve whole number problems involving addition and subtraction".

This approach will be termed *strong* criterion-referencing because of the strength of the descriptive inferences about pupils' attainments which it purports to make possible. The psychological and epistemological naïvety of strong criterion-referencing has been discussed in detail elsewhere (Cresswell and Houston, 1991). Wolf has also offered some trenchant

criticisms (Wolf, 1993) and, at the time of writing (1996), the approach is increasingly being questioned. Nonetheless, it is worth considering in a little detail here, if only to provide additional warnings to those who might be tempted by it in the future.

In the terms of the DIE model, strong criterion-referencing represents an attempt to make assessments which are genuinely deserving of the term *descriptive*. This has involved the replacement of numerical scores with verbal descriptions. Thus, instead of awarding, say, 15-20 marks for demonstrating a firm grasp of all aspects of Mr Grimes' character, the marking process would simply state that the pupil *demonstrated a firm grasp of all aspects of Mr Grimes' character*. The problem then, of course, is the combination of the extremely fine level of detail involved in such descriptions and the difficulty of aggregating verbal descriptions from, say, more than one assessment task.

To illustrate this point, consider how the attainment of a pupil who *demonstrated a firm grasp of all aspects* of Mr Grimes' character but only *demonstrated a knowledge of the main aspects* of Tom's character can best be described. The usual response to this problem is to attempt some verbal synthesis such as *demonstrated some understanding of two of the book's main characters*. The use of phrases like *some understanding* in this type of synthesis is at the heart of the strong criterion-referencing approach but, unfortunately, involves an unrealistic view of the precision of meaning which can be achieved in a short piece of natural language. The formulation *some understanding* is an attempt to define a state somewhere between *a firm grasp of all aspects* and *a knowledge of the main aspects* but whether or not this particular formulation is a good one is beside the point here. The fundamental question is whether such a verbal averaging process can ever result in a more precise description than a numerical one: is *demonstrated some understanding of two of the book's main characters* any more informative than *scored 65% on questions about two of the book's main characters*? Despite the apparent precision of the numerical statement, neither it, nor the verbal synthesis, conveys specific information about the pupil's attainment. Sadler (1987) argues cogently that this is inevitable.

Thus, once descriptive verbal statements are combined into a summarising synthesis, the result is no more informative than an aggregated numerical score and, consequently, strong criterion-referencing fails to offer descriptive reports which are any more informative than those of conventional criterion-referencing. The same problem with the result of verbal synthesis occurs at every point in the assessment hierarchy where aggregation occurs (see Figure 3.2) and this has led to attempts to devise aggregation rules for descriptions of pupils' performances on individual tasks (or sets of tasks) which preserve specific descriptive information in the resulting, more general summary. The theoretical issues raised by such rules are discussed in Chapter 4. The results of an attempt to apply them are described in Chapter 8.

Another aspect of the verbal synthesis proposed above is also worth discussing: the conflation of questions about Mr Grimes and Tom into questions about *two of the book's main characters.* This raises questions of interpretation, as distinct from description, because it invites the inference that the pupil can demonstrate some understanding of **any** two of **any** book's main characters. In fact, as the DIE model makes clear, the validity of any such inference depends crucially upon the relationship between the characters and characterisation in *The Water Babies* and the assessment domain as a whole.

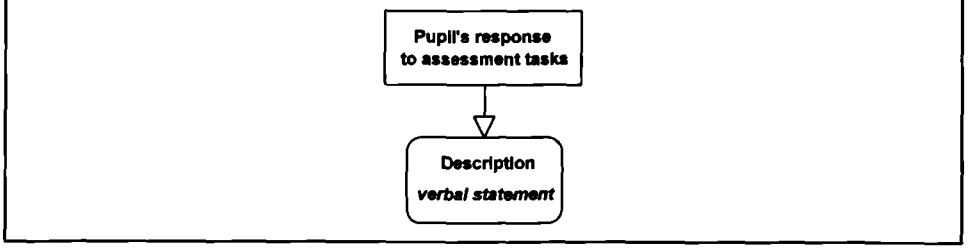
A considerable practical problem with strong criterion-referencing is the quality of the evidence which is used to identify verbal statements of attainment with individual pupils' attainments. In contrast with the usual criterion-referencing approach, some of the recent (1991 to 1994) approaches to assessment of the National Curriculum assume that adequate evidence can be provided by a very small number of assessment tasks (see, for example, Ruddock, *et al*, 1993). Moreover, the basic approach can be characterised as opportunistic in that control of the assessment tasks used to provide evidence about any particular pupil is not given a great deal of importance beyond an acknowledged need for superficial relevance to the statement of attainment concerned. This approach ignores both recent research and the experience gained over at least half a century by test constructors and examiners.

As Foxman *et al* (1985) have shown in great detail for Mathematics and Pollitt *et al* (1985) have shown for a wider range of school subjects, the details of a task, even apparently superficial details, affect pupils' responses considerably. It is hardly surprising, in these circumstances, that experienced examiners are unable to construct examination questions of specified difficulty (Good and Cresswell, 1988a); that it should be a commonplace among test constructors that item difficulties are unpredictable (eg Wesman, 1971) or that two tests constructed to the same specification will not be equally difficult or have identical score distributions (Braun and Holland, 1982).

There is a clear contrast between some recent high-profile developments in the UK and normal practice in criterion-referenced assessment where great pains are taken to control the representativeness of the set of assessment tasks used. It is axiomatic that, without information about the representativeness of the assessment tasks, general interpretations of decriptive reports cannot be made. Moreover, the interpretative problem is not reduced by the often recommended practice (for example, see Sadler, 1987) of quoting examples of questions to which the descriptions relate. Unfortunately, as illustrated by the synthesis *two of the book's main characters* which was mentioned earlier, under strong criterion-referencing the need for interpretation can be obscured by the superficially general relevance of the verbal descriptions which it produces.

In keeping with this analysis, Figure 3.5 illustrates the limited nature of strong criterion-referencing as it has been practised in the assessment of the National Curriculum in England and Wales.

Figure 3.5
Strong criterion-referencing in terms of the DIE model



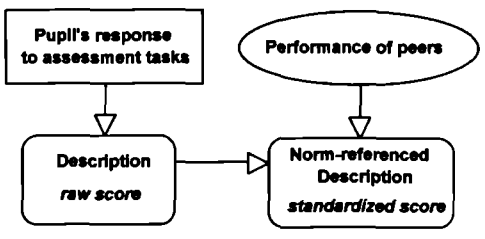
3.2.3 Conventional Norm-referencing

Like criterion-referencing, the term *norm-referencing* has recently come to be used to describe approaches other than the original one for which it was coined. In this section, norm-referencing is used in its original sense of *standardizing* - locating each pupil's score within the distribution of attainment of his or her peers. The other, more recent, usage of the term *norm-referencing* is covered in the following section.

One of the powerful features of the DIE model is the understanding which it gives of the relationship between conventional norm and criterion-referencing. Whereas criterion-referenced tests may produce either interpretative or evaluative reports, conventional standardized tests usually produce only descriptive ones. This is not, however, due to their norm-referenced nature itself, but is a consequence of the accompanying approach conventionally taken to the definition of, and sampling from, the domain being assessed.

Conventionally, norm-referenced (standardized) tests do not have tightly defined assessment domains and their representativeness, therefore, is difficult to establish. It can be argued that, as a result, such tests do not provide more than descriptive reports of pupils' attainment, interpretation and evaluation being left to the users. (Perhaps this is the fundamental source of the criticism which is sometimes made of norm-referenced tests that they assess only what is in the test.) In terms of the DIE model, the process of norm-referencing a conventional standardized test can be seen to be nothing more than a transformation of the original score scale of the descriptive report into another form. This is illustrated by the modified and reduced DIE model shown in Figure 3.6.

Figure 3.6
Conventional norm-referenced test in terms of the DIE model



A particularly interesting feature of such norm-referenced tests is that, by reporting in terms of the distribution of peer attainment, they invite evaluative judgements to be made. The evaluative leap to satisfaction with a score above, say, that of 90% of a pupil's peers is difficult to resist even though, without knowledge of how the test represents an assessemnt domain of interest, satisfaction is unjustified. On the basis of the present analysis, evaluative comparisons between pupils who have taken the same test are the only theoretically legitimate use for conventional standardized tests.

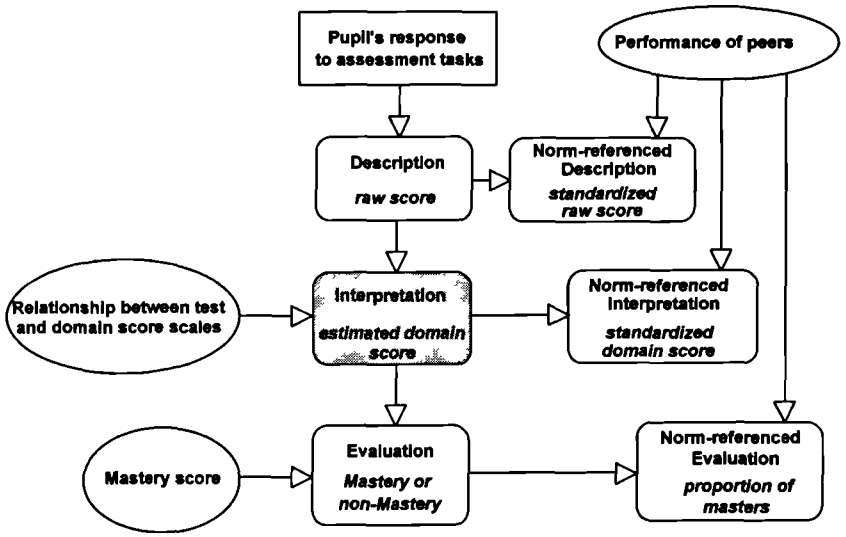
Although it is conventional to norm-reference descriptive reports, the DIE model makes it clear that interpretative reports or evaluative reports could equally well be norm-referenced. Thus, there is no inherent contradiction in a test being **both** criterion- and norm-referenced. Indeed for many purposes, interpretative and evaluative reports from criterion-referenced tests would be enhanced by being accompanied by norm-referencing information. For example, a teacher who knew that the proportion of pupils in his or her class who had achieved "mastery" of a particular domain was substantially below the peer average, would, at the very least, be likely to interpret the criterion-referenced information more critically. It is hard to argue that the search for explanations for such a state of affairs would not bring positive benefits to that teacher's understanding of his or her task and the context within which it was being carried out. Figure 3.7 illustrates, in terms of the DIE model, the relationship which exists between norm- and criterion-referencing.

3.2.4 Public Examination Awarding

As with criterion-referencing, the term norm-referencing recently acquired a new usage in the UK. Paralleling the development of the notion of strong criterion-referencing, norm-referencing came to be applied to any assessment system which used numerical marking to determine pupils' results. In particular, the traditional marking and awarding procedures of the public examining boards have been erroneously called norm-referencing. There is considerable irony in this because, with the perspective provided by the DIE model, current public examining procedures can be seen to have more in common with criterion-referencing than with norm-referencing. In particular, as Christie and Forrest (1981) argued, for public

examinations there is usually a clearly enunciated assessment domain and, in the question setting process, considerable effort is made to ensure that the examination as a whole represents it effectively. Information about the sampling frame used to construct the papers is frequently included in the syllabus in the form of an assessment grid.

Figure 3.7
DIE model of a conventional criterion-referenced mastery test showing norm-referencing opportunities



Nonetheless, public examination syllabuses do not define the assessment domain of the examination with the precision required by conventional criterion-referencing and the representativeness of the question papers is established in terms of wide areas of knowledge and generally defined skills. This reflects the breadth of the domains which public examinations assess and it is this aspect of them which distinguishes them from conventional criterion-referenced testing. In terms of the DIE model, public examinations can be represented in outline as shown in Figure 3.8. The structural similarity between this figure and that for criterion-referenced mastery testing (Figure 3.4) illustrates the point made in the previous paragraph.

Figure 3.8
Public examining in terms of the DIE model

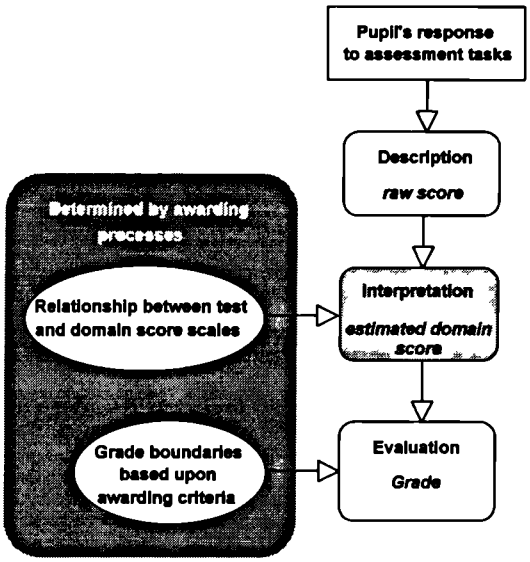


Figure 3.8 also identifies the matters which are considered during the awarding processes which are the subject of the present study. Awarding procedures were described, in general, in Chapter 1 and it is apparent that the interpretative domain score is never explicitly formed for public examinations. The need for awarding nonetheless to consider the relationship by which the domain score **could** be formed is one of the key issues addressed in this study and is considered in more detail in Section 3.5.2, below.

It is worth, at this point, asking why public examinations report in evaluative terms, given their primary purpose as providers of information for selection (see Chapter 2). In the most common cases of selection for employment or for entry onto educational courses, the candidates being selected will have been assessed using different examinations on different occasions. The information required by selectors cannot, therefore, be the descriptive raw scores of each candidate because raw mark scales are essentially arbitrary, being dependent upon the particular assessment tasks set in each examination and on each occasion. As a result, comparisons between raw scores from different examinations are not valid.

However, within a given assessment domain, selectors could use interpretative reports to choose between candidates. Unfortunately, practical selection decisions frequently involve

comparing candidates whose public examination results are based on assessment in quite different school subjects (see Chapter 2) and even within the same subject, the examinations of the different boards differ in the details of their assessment domains. However, domain scores are not directly comparable across different domains and there are major theoretical problems and significant practical difficulties which make it extremely difficult for individual selectors to make fair comparisons of scores relating to different domains. Evaluative reports, in the form of grades, are therefore provided by the examination system, with a given grade claimed to represent a comparable standard of attainment from any examination in any assessment domain. The theoretical problems of this claim and the sense in which it can be substantiated are considered in detail in Section 3.4, below.

3.3 AWARDING AS AN EVALUATIVE ACTIVITY

3.3.1 Previous work on Grade Criteria

If awarding is seen as an evaluative process, it appears reasonable (as Christie and Forrest, 1981, argued) to try to identify the criteria which awarders use to decide the relative merits of, and thus the grades awarded to, candidates' work. It is also argued (see, for example, DES, 1982) that if these criteria could be identified, they could be written down and used to communicate the standards required for each grade to candidates, teachers and selectors. However, it is important to note the difference between the previous two sentences in the way in which they use the word *criteria* (see Section 3.2.1). Christie and Forrest conceptualise a grade criterion as an attribute to be assessed; the DES paper sees it as a standard. In the early 1980s, there was a considerable amount of work done in Britain on this ambiguous notion of *grade criteria* for public examinations (Hadfield, 1980 [quoted in Christie and Forrest, 1981]; Orr and Nuttall, 1983; SEC, 1984; Forrest and Orr, 1984; Orr and Forrest, 1984; Bardell *et al* 1984; SEC 1985b; Long, 1985; SEC, 1986; SEC,1987). Almost all of this work was based on the view that grade criteria should be standards; that is, written statements which prescribe the level of attainment required to justify the award of a particular grade. It is worth noting that by 1984 Forrest and his co-workers also followed this line, explicitly confusing the two conceptions of criterion in the following definition: "[criterion-referencing is]

relating the award of grades to *specified levels of attainment* in *defined aspects of the subject*" (italics added).

These various attempts to develop explicit standard-setting grade criteria concerned themselves almost exclusively with apparently observable qualities in candidates' work. This reflected an implicit underlying theoretical model which viewed the task of awarders simply as the identification of appropriate qualities in candidates' work, rather than the formation of a reasoned evaluation, based upon such qualities. The effects of this theoretical error became manifest in the work done by all the examining boards at the request of the School Examinations Council (SEC, 1986; SEC, 1987). Although it was possible to write standard-setting grade criteria (either *ab initio* as in the original SEC work or as a result of perusing scripts, as in later work), in use, they proved not to apply to some candidates' performances which were awarded the grades in question by conventional procedures (Cresswell, 1987c). Nor was this the result of errors in awarding grades in the first place, since the examiners involved in the studies characteristically avowed the appropriateness of the grades concerned. Similar results were obtained in the work carried out by Forrest and his co-workers (Forrest and Orr, 1984; Orr and Forrest, 1984; Bardell *et al* 1984) who, in their general conclusions, published in the reports of all their studies, observed:

> "performances in existing examinations that would result in the award of particular grades may not qualify for those grades if the criteria considered relevant were applied."

Thus, despite considerable efforts, no system of standard-setting grade criteria for awarding public examinations has been made to work successfully. The studies cited above were, nonetheless, instrumental in the British development of the ideas of strong criterion-referencing which were discussed in Section 3.2.2; indeed, the use of grade criteria which purport to define standards is vulnerable to exactly the same arguments relating to psychological and epistemological näivety (see Cresswell, 1987b and 1987c).

A major problem for the development of standard-setting criteria lies in the multi-attribute nature of the attainments being assessed and the inability of the criteria to specify the weight which should be attached to each attribute when judging individual candidates' work. Wilmut

and Rose (1989), who had been trying to use descriptive statements akin to grade criteria to set standards for the award of "levels" (simply numerically denoted grades) in a modular assessment scheme summarised the position as follows:

> "The real difficulty ... has been the breadth of these descriptions; because they relate to a whole module they must encompass a wide range of attributes, and it is very difficult to characterise these adequately in statements which are brief enough to be readily usable for decision-making. There is, of course, no difficulty in summarising the descriptions in order to convey the 'flavour' of a level, but reduced statements of that kind are of little use for deciding about the worth of students' work. In practice, such work is often in several pieces, each of which has to be assessed separately, and is usually intended to exhibit several connected but separate attributes. Thus we may, for example, be attempting to assess a piece of project or fieldwork under the general objectives of
>
> seek out information
> use that information
> communicate the results.
>
> Level descriptions have adequately to cope with these three objectives, properly indicating their relative importance, and be expressed in a way which allows us to make sensible decisions about pieces of work in which the objectives are met to different degrees. Thus, one student might be good at seeking out and using but poor at communication, whereas another may use information less well, but get the communication about right. Both might merit a certain level, and the descriptions must enable us to see that this ought to be so."

As far as it goes, Wilmut and Rose's diagnosis of the problem is correct but they still appear to believe that practical standard-setting criteria can be written which reduce the formation of evaluative judgements to the simple identification of objective qualities. In fact, as the next section demonstrates, the failure of every known attempt to construct and use such grade criteria is not a failure to do the job properly by those who have tried. It is an inevitable consequence of the fundamental nature of evaluative reasoning.

3.3.2 The reliability and nature of evaluative reasoning

If, as is argued here, grade awarding is not a matter of identifying performances which meet some set of objective grade criteria but is a subjective process more akin to evaluating a work of art, immediate questions arise about its acceptability. Given the importance to individuals of the selection decisions which rest upon examination results, the issue of subjectivity is clearly important. A key requirement is that those making judgements of others' attainment must be generally accepted as competent and trustworthy. The implications of this are discussed in

detail in Section 3.4.2.1 and again in Chapter 9. Assuming, for the moment, that the requirement of competence and trust is met, it remains important to clarify a further aspect of the subjectivity involved in examination awarding. Although the value judgements of awarders cannot be facts amenable to empirical verification or epistemological justification, they can, nonetheless, be the results of a rational process and be supported by reasons (Fogelin, 1967; Beardsley, 1981) if not pure deductive or inductive reasoning (Best, 1985). Value judgements based upon reasoned argument are not simply emotional or intuitive responses and should not, therefore be assumed to be necessarily capricious or unreliable in the sense of being difficult to replicate.

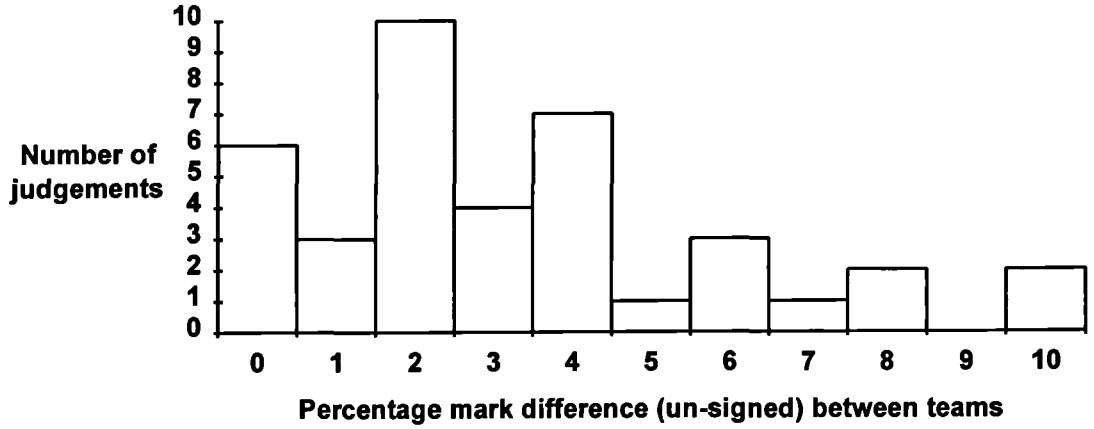
In fact, as noted in Chapter 2, there appears to be a tolerable degree of reliability in the value judgements made for grade awarding purposes. Indirect general evidence for this comes in the form of the results of the many cross moderation comparability studies which have been carried out over the years (Forrest and Shoesmith, 1985) and have implied that the value judgements of different groups of awarders agree reasonably well, most of the time. Little direct evidence of the reliability of awarding decisions is available. However, some was collected by Good and Cresswell (1988a) who replicated the awarding meetings for their experimental examinations in French, History and Physics. Good and Cresswell concluded that

> "...different groups of grade awarders can reach decisions about final grade boundaries which are sufficiently similar to be acceptable, given the inherent imprecision of the examining process."

This conclusion was formed on the basis of the consequences for candidates' final subject grades after component boundary judgements had been combined and some cancellation of component grading errors had occurred. The final qualification in Good and Cresswell's conclusion is also important. The percentage of candidates whose subject grade changed if one awarding team's boundaries were substituted for another's was 13% in French, 17% in Physics and 38% in History. Good and Cresswell point out that, although large, such changes are unexceptional in the context of the grade differences which occur between markers, even with very high levels of marking reliability. (Good and Cresswell refer to work by Wilmut (1981) which shows that, for the GCSE grade scale, 38% of candidates changing grade

corresponds, approximately, to an inter-marker reliability coefficient of 0.96.) A pooling of Good and Cresswell's (1988a, Pages 21-23) data for all three subjects, shows that the disagreements between different teams of awarders on 39 component grade boundary decisions were distributed as shown in Figure 3.9.

Figure 3.9
Distribution of differences, expressed as percentages of the available marks, between different teams of awarders independently making 39 grade boundary judgements



Although the mean mark difference between the judgements of the teams is only 3.3%, further work on the reliability of awarding judgements would clearly be desirable to illuminate why some individual judgements show atypically large disagreement between different teams of awarders. Nonetheless, quite good agreement seems to be obtained for most awarding judgements.

That most awarding judgements are reasonably replicable would not surprise authors such as Fogelin (1967) and Best (1985) who have made a study of the processes by which value judgements are formed. Although the process of evaluation is not one of deductive or inductive logic, it is nonetheless rational. Reasons can be adduced for evaluative judgements and these reasons can be assessed by well accepted criteria. In particular, those reasons which are matters of fact can be verified and combinations of reasons can be assessed for consistency, one with another (Collingridge, 1982; Best, 1985). Thus, although there is no

way of **proving** the accuracy of value judgements, it is not true that rational debate cannot take place about them, nor that discussion cannot usefully clarify the reasons for such judgements and thereby persuade other competent judges of the accuracy of a particular judgement.

Nonetheless, it remains the case that two awarders might agree completely about the relevant reasons for their judgements but still differ in that judgement. This is because, understood in terms of deductive logic, value judgements can be seen as following from a set of premises (the agreed reasons) and at least one further prescriptive premise (Fogelin, 1967). In this way, Fogelin argues that value judgements are warranted prescriptions which recommend a course of action. Whether or not this is true in general (see Pole, 1961), it appears true in the case of examination grades. The award of a Grade B, for example, amounts to the prescription to select this candidate in preference to any other one with a worse grade but not to select this candidate in preference to one with a Grade A; the warrant for this prescription being grounded in the description of the candidate's attainment.

Clearly, the adoption of different prescriptive premises would be sufficient to explain irreconcilable disagreement between two awarders who, nonetheless, shared the same view of the qualities of a candidate's work. Thus, if awarders are necessarily to agree about value, given that they agree about reasons, they must share the same prescriptive premise(s). Billington (1988) argues that prescriptive premises can also be discussed rationally, often in terms of the value of their consequences in other cases. However, there is a circularity here in the appeal to other value judgements to justify prescriptive premises and the consequence is that awarders are most likely to reach agreement after discussion of their respective views of candidates' work, if they share a common understanding about the prescriptive import of those qualities.

Sadler (1985, 1987, 1989) has written extensively about the way in which educational evaluation involves the use of *tacit standards* held by a *guild of professionals.* It is this aspect of the awarding process which enables awarders, at least in theory, to progress towards agreement about value judgements by discussing their reasons and prescriptive premises.

The standards remain *tacit* because they are internalised from lengthy experience of pupils' attainments and the ways in which these are rewarded in examinations, modified by working alongside other members of the *guild*. In this connection, it is noteworthy that, whereas Sadler's 1987 paper advocates the use of written standards, supported by examples, to communicate the guild standards to others, his later 1989 paper recommends that the best way to teach students how to evaluate their own work is to give them direct, guided experience in evaluation which involves discussion of the criteria used in many specific instances.

This approach is also consistent with Sadler's other main contention: that the criteria which provide appropriate reasons for an educational value judgement differ according to the nature of the work being evaluated. This has long been recognised as a feature of judgements of aesthetic value (see, for example, Aldrich, 1963). Indeed, the presence of the same quality may be reason to value one object highly but not another (Pole, 1961) because of other features of the second object which change the value attached to the quality in question. Work on the psychology of evaluative judgement has similar implications. Eiser (1990) argues that a person presented with new information searches for relevant conceptual categories with which to encode it, starting with those immediately accessible in memory and continuing until relevant ones are found. Here are the fundamental reasons why the application of concise sets of written criteria does not replicate value judgements made by suitably qualified judges and they doom the development of explicit grade criteria for public examinations to failure.

Note, however, that this is not to say that concise statements cannot be constructed which, in Wilmut and Rose's (1989) terms, "convey the flavour" of a grade. Such *grade descriptions* have been included in GCSE syllabuses since 1988. However, they describe a paradigmatic attainment worthy of the grade, rather than the attainment of every candidate awarded the grade and cannot, for the reasons discussed above, be used as criteria for judging the attainment of all candidates.

It is worth noting that the theoretical analysis in this section is consistent with the normal public examining practice of separating the function of marking candidates' work from that of

awarding. Although marking has an evaluative aspect (see Section 3.1) in many subjects, it is carried out by means of an analytical mark scheme which specifies precisely which criteria are assessed and, through numerical mark values, indicates how strengths and weaknesses in different criteria are traded off against each other. This approach is usually justified in the interests of forming a reliable rank-ordering of large numbers of candidates when many different examiners are, for practical reasons, involved in marking the candidates' scripts.

If marking were to be replaced with the direct evaluation of every candidate's work, the same large number of examiners would all need to share the same tacit standards. For practical reasons, these examiners would also be denied access to the reasoned argument which, it was argued above, assists the formation of consistent value judgements. In addition, when responses to the same set of assessment tasks are being assessed, the use of a set of pre-determined criteria with pre-determined trade-offs is facilitated because the range of evaluatively relevant qualities is limited by the assessment tasks. (Note, in passing, that it is consistent with this analysis that assessment tasks such as essay writing, which admit of a wide range of responses, tend to be those where evaluation plays a larger part in the marking process and reliability is lower - see Newton, 1996.)

The range of possibly relevant assessment criteria and the tendency for them to be differentially relevant to evaluation is very much greater if responses to different assessment tasks must be comparably evaluated. This must be done, for example, if one year's examination grades are to be comparable with another's and awarding meetings are held precisely with the intention of producing comparable grades from different examinations.

3.4 COMPARABILITY OF STANDARDS

The claim made by public examinations that a given grade represents the same standard of attainment, regardless of the subject or examination from which it comes, is a very strong one which is a key justification for the use of public examination results in selection. (It is also a premise of recent legal requirements upon schools to publish their public examination results as indicators of their effectiveness.) Indeed, to argue the need for comparable examination

grades is simply to re-state the need for selections made on the basis of them to be fair in terms of the meritocratic philosophy which underpins their use (Chapter 2). What constitute meritocratically fair selections? In Chapter 2, it was argued that there is a strong strand in meritocracy, borne out by the behaviour of selectors, to the effect that individuals should be selected on the basis of their current attainment. Since selections are made from among candidates who have taken different examinations, this has led to a perceived need to establish quantitatively equivalent levels of attainment across qualitatively different assessment domains. This is formally impossible in terms of the theory of educational measurement (see, for example, Goldstein, 1986b). Comparability is not, however, a purely technical matter and can only be fully understood in relation to the evaluative selection processes in which examination grades are used. Below, in Section 3.4.2, a new definition of comparability is proposed which gives appropriate emphasis to the social nature of the selective function of public examinations. First, however, the traditional approach to the topic is discussed.

Because of the manifest importance of comparability for the fairness of selection, there have been many studies attempting to establish whether or not the grades from particular public examinations reflect comparable standards. The GCE examining boards in England and Wales published two summaries of the 51 such studies which had been carried out up to 1985 (Bardell, *et al,* 1978; Forrest and Shoesmith, 1985). The Schools Council published a series of major reports in the 1970s (see Willmott, 1980) and since 1985 a further 22 comparability studies have been carried out for GCSE and A-level examinations with the reports being published by, and available from, the examining boards and groups. These studies cannot be reviewed in detail here but the methodologies which have been adopted are of interest. They can be divided into two main types: those which involved statistical analysis of the distributions of candidates' grades (or equivalent analyses) and those (called cross-moderation studies) which depended upon examiners' judgements to identify comparable work from the examinations being studied. Some of the studies used both approaches.

Although rarely explicit about their theoretical perspective, those responsible for comparability studies which analyse the distributions of candidates' grades seem generally to subscribe to

the erroneous view that comparability is a technical problem with a technical solution within measurement theory. The studies using examiners' judgements have also been largely atheoretical, simply taking it as reasonable that examiners are able to judge comparable standards of attainment on different assessment domains. Where they have an identifiable theoretical base, however, it appears to be to treat examiners' judgements as a criterion variable, assuming that scripts awarded the same average judgement should, on average, be awarded the same grade in comparable examinations. As a result, some such studies (for example, Houston, 1980) have gone to considerable lengths to obtain agreement among the examiners involved about the judgemental criteria which they use. This conception of the cross-moderation methodology also identifies comparability as a purely technical problem of measurement theory.

Since GCSE examinations were introduced in 1988, some subjects have been examined using differentiated papers. In these examinations, candidates taking different combinations of papers, which are designed to differ in difficulty, are awarded a grade on a common scale. The problems of setting comparable standards on the different combinations of papers used in such examinations were studied in depth by Good and Cresswell (1988a and 1988b).

3.4.1 Comparable standards defined statistically

What has *comparable*, as applied to examination grade standards, normally been taken to mean in previous statistically based studies? In general, it has not been taken to mean that an **individual** taking two comparable examinations should necessarily be awarded the same grade. Not only are the assessments involved subject to various sources of error which would prevent this outcome from routinely occurring, but it is also accepted that individuals attain differently in different assessment domains. Thus, the notion of comparability applied to public examination results is concerned with groups of candidates. For example, talking about the particular case of comparability between examining boards (and assuming a great deal about other variables relevant to performance in public examinations which will be considered shortly) the Forum on Comparability set up by the *Schools Council* (which had a general

responsibility for British public examinations until the mid 1980s) said:

> ... "the expectation is that had a group of examinees followed another board's syllabus and taken its examination, they might reasonably be expected to have obtained the same average grade." (Schools Council, 1979)

The definition of comparability implicit in this quotation can usefully be elaborated to cover the comparability of every grade, by replacing the reference to the "average grade" by a reference to the **distribution** of grades for the group of candidates. On this basis, a variety of definitions of comparable examination standards can be explored by considering some of the different ways in which differences between grade distributions have been analysed in previous studies.

Clearly, for analysis of examination grade distributions to provide a reliable indication of the comparability of grade standards, the grade distributions need to depend only upon those standards. In fact, of course, grade distributions also reflect the attainments of the candidates who take the examinations and these attainments are the result of the interaction of many different variables. Systematic differences between the self-selected groups of candidates who take different examinations are to be expected and the various statistical techniques which have been used for investigating examination comparability by analysing grade distributions attempt to control for differences in the attainment of the candidates taking the examinations in one of two ways. Either an independent measure of attainment is employed or indirect control of attainment is attempted through school and student variables which influence it. Once this has been done, it is argued, any remaining differences between grade distributions indicate lack of comparability between the examinations.

Three major issues immediately arise, however. First, candidate attainment is, itself, affected by features of the examinations and their syllabuses (examination variables) as well as characteristics of the candidates' schools (school variables) and characteristics of the candidates as individuals (student variables). As a result, controlling for differences between different groups of candidates in terms of attainment will, in passing but unavoidably, also

control to some extent for the relevant features of the examinations and their syllabuses. These features include the demands of the skills and subject content specified in the syllabus and its organisational aspects. Syllabuses can differ in both these respects and both may affect the motivation of pupils and the ease with which they are taught and learn. For example, at one extreme, a syllabus can be presented as a simple list of skills and content; at the other extreme, the same skills and content can be usefully structured and accompanied by examples and guidance for teachers. The question of importance in the present context is whether a syllabus (and associated examination) which makes the subject more accessible or more motivating and thus produces a grade distribution which is skewed towards the better grades, therefore effectively sets lower standards for the award of grades than an obscure syllabus which makes learning difficult. Questions of this type must be answered before any statistical analysis of grade comparability can be interpreted.

The key to the answer lies in the function of the examinations. If this is the provision of information for meritocratically fair selection processes in which current attainment is the selection criterion, then fairness requires that those who study syllabuses which organisationally facilitate learning should have a greater chance of being selected than those who study less enabling syllabuses. It follows that they should, in their examination grades, be rewarded for their higher attainment even though it is a consequence of the organisation of the particular syllabus which they have followed. The corollary of this is that different grade distributions from two examinations which differ only in the organisational aspects of their syllabuses do not necessarily indicate a lack of comparability in the grading standards being applied.

On the other hand, if two syllabuses are judged to be of equal value but nonetheless differ slightly in terms of the intellectual demand of their content, consequential differences between their grade distributions would imply a lack of comparability because it is meritocratically unfair for a candidate to have less chance of being selected because he or she showed poorer attainment on a harder assessment domain. (Indeed, it is concern about differences of this sort which motivates the demand for comparable grade standards in the first place.) However, differences in syllabus demand are also likely to affect the motivation of pupils and

differentially to facilitate teaching and it was argued in the previous paragraph that motivational differences between syllabuses are a legitimate reason for differences in their grade distributions.

Because of interactions of this type, in practice it is impossible to distinguish between the effects of differences in the intellectual demands of syllabuses and the effects of differences between their organisational aspects. As Goldstein (1986b) points out, comparability studies based upon the statistical analysis of grade distributions are forced to ignore such differences and generally take as a working assumption that the effects of syllabuses and examinations upon teaching, learning, examination entry policy and so on are identical for all the examinations studied.

The second theoretical problem with definitions of comparable standards which depend only upon identical grade distributions concerns identifiable subgroups of candidates. Even when two grade distributions are identical for the whole groups of pupils whose attainment they describe, they are frequently not identical for well-defined subgroups within those groups. For example, the differences between boys' and girls' performances in GCSE examinations are well known and depend, at least in part, on the assessment techniques used (see, for example, Murphy, 1982b; Cresswell, 1991). Thus, if two examinations using different techniques have identical grade distributions for boys and girls combined, they will not necessarily have identical grade distributions for the boys and girls considered separately. It is unclear how this can be accommodated if the test of comparable standards is identity of score distributions, since the meaning of comparability is unclear if it is limited only to a particular subgroup of candidates. Certainly, comparable standards defined only for a particular subgroup of candidates are insufficient to permit the general use of examination results in selection. Only if the control of the other variables which determine grade distributions also removes any observed differences between well-defined subgroups of candidates, can the difficulties raised in this paragraph be avoided.

The third, and most fundamental, problem with purely statistical methods of studying comparability concerns the content of the syllabuses. Christie and Forrest (1981) pointed out

that examination standards can only be comparable if the syllabuses concerned define assessment domains which are appropriate to the "particular subject at a particular level of education". As a thought experiment to illustrate this point, consider a GCE A-level Physics syllabus which was entirely descriptive and did not contain any mathematical treatment of the phenomena covered. An examination could undoubtedly be set on such a syllabus which would produce the same grade distribution for a given group of candidates as they achieved on a more conventional Physics syllabus. Would, therefore, the grade standards of the two examinations be comparable? The answer to this question clearly depends upon the value given to a non-mathematical Physics syllabus at A-level. Thus, defining comparability of examination grading standards solely in terms of identical grade distributions is inadequate unless it can be assumed that the syllabuses upon which the examinations are based define assessment domains of equal value. This assumption is therefore crucial to all statistical approaches to setting comparable examination standards but, for the sake of the discussion, it will not be challenged immediately. In Section 3.4.2, however, the vital importance of the valuation of the assessment domain is re-introduced and a new theoretical basis for the definition, setting and maintenance of standards is proposed which accomodates it.

As noted earlier, a striking feature of most of the statistical work on examination comparability which has been carried out is the lack of an explicit theoretical basis. In particular, few published statistical studies, from the Schools Council work in the 1970s (Willmott, 1980) to the recent work by Fitz-Gibbon and Vincent (1994), have attempted to test the three crucial assumptions set out above: that different syllabuses have identical effects upon motivation, teaching, learning, school entry policy and so on; that the observed relationships between the grade distributions are the same for identifiable subgroups of candidates; and that the value of what is assessed in the examinations studied is comparable. Instead, those responsible for comparability studies which analyse the distributions of candidates' grades seem, relatively uncritically, to have adopted one of the variety of implicit definitions of it which are explored below.

3.4.1.1 The no-nonsense definition

> *Two examinations have comparable standards if the distributions of grades which they produce are identical.*

From the foregoing discussion, it will be clear that this definition is inadequate. It assumes that in terms of relevant student variables (ability, prior achievement, motivation) and relevant school variables (general effects, effectiveness of subject teaching, entry policy) together with the effects upon these of the two syllabuses and examinations, the groups of candidates entering the two examinations are identical. Most of these assumptions are known not to hold in practice. In addition, comparability defined this way will not always hold for subgroups of candidates considered separately even if it does for the group as a whole. Nonetheless, the no-nonsense definition of comparability is worth mentioning because it is frequently used in press discussion of examination results. It has also been used by British government advisory bodies as the basis for querying, with the examining boards and groups, the comparability of particular examinations and has been used during the deliberations of the *Independent Appeals Authority for School Examinations* (see, for example, IAASE, 1993) to which candidates can appeal if they believe their public examination results to be in error.

3.4.1.2 The same-candidates definition

> *Two examinations have comparable standards if, when the same group of candidates is entered for them both, the distributions of grades which they produce are identical.*

This definition assumes that motivation, prior achievement and the influence of relevant school variables (effectiveness of subject teaching, entry policy), together with the effects upon these of the two syllabuses and examinations, are identical when the same candidates tackle two different syllabuses and examinations. There is no reason why these assumptions should hold in practice simply because the same students are involved. Nonetheless, the intuitive appeal of this definition of comparable standards is considerable; it is used by teachers who enter candidates for examinations in the same subject with different boards and expect to get identically distributed results. This definition is also the basis for a particular approach to comparability between examinations in different subjects, known as *subject pairs analysis*, which has long been controversial on account of the assumptions just outlined (see Forrest and Shoesmith, 1985).

It is worth briefly mentioning one interesting condition under which the assumptions of the same-candidates definition might be thought of as axioms, rather than challengeable assumptions. If all the pupils in a particular age cohort take two examinations in different subjects, then it appears reasonable to define comparable standards between the two subjects concerned in terms of identical grade distributions from the two examinations. This situation is most closely approached in the public examination system by GCSE examinations in English and Mathematics. The argument is as follows: it is difficult to see what meaning can be attached to the notion that the sum total of all the English teaching in the educational system is more or less effective than the sum total of all the Mathematics teaching. Similarly, if there is, on average across all those who study them, a motivational difference between the two subjects, is this not most parsimoniously treated as simply characteristic of learning in the two subjects?

Fundamentally the same argument can be applied to all the variables upon which the attainment of the pupils in the two subjects depends. If it is accepted, it follows that, on condition that the examinations are taken by the entire age cohort, the same-candidates definition of comparable standards is theoretically coherent and might be useful. However, the implications of differential comparability for different candidate subgroups would need to be dealt with if it occurred and there would remain a significant problem if the relationship between performances in the two subjects changed over time. In these circumstances, either comparability between subjects or across time could be maintained, but not both. Note that although the age cohort condition is not generally met in public examining, it is quite closely met by the statutory assessments at other ages of the National Curriculum.

3.4.1.3 The value-added definition

> *Two examinations have comparable standards if two groups of candidates with the same distributions of ability and prior achievement receive grades which are identically distributed after studying their respective syllabuses and taking their examinations.*

This definition assumes that, in terms of other relevant student variables such as motivation and relevant school variables (general effects, effectiveness of subject teaching, entry policy) together with the effects upon these of the two syllabuses and examinations, the two groups of candidates are identical. Most of these assumptions are known not to hold in practice, so

analyses which simply allow for ability and prior achievement are unlikely to give reliable information about comparability. However, if the prior achievement in question is the result of schooling in the institutions which enter the candidates, allowing for it may also partially, but to an unknown extent, allow for the effects of some school variables. This definition has frequently been used; most recently by Tymms and Fitz-Gibbon (1990), Fitz-Gibbon and Vincent (1994) and Tymms and Vincent (1995), but most extensively by the Schools Council researchers in the 1970s (Willmott, 1980) who sometimes used a specially written reference ability test, known as *Test 100*, as a statistical control.

Because this definition has been extensively used in the past and is now enjoying something of a comeback, it is worth considering its underlying rationale from another angle. In particular, although Fitz-Gibbon, Tymms and Vincent do not discuss the theoretical basis of their studies at all, there does exist a theoretical perspective which appears to make the assumptions listed in the previous paragraph unnecessary. This perspective starts from the premise that it is ability or aptitude which should be the basis for selection, rather than attainment, and treats differences in measured attainment from subject-specific examinations as uninformative "error" caused by the vagaries of schooling and the examining process. Differences in measured comparability between candidate subgroups is seen as evidence of bias in the assessments. There is some evidence that the thinking behind the Schools Council reference test comparability studies followed this line; Nuttall and Willmott in 1972 suggested that there was a case worth considering for using "a single general intelligence test" in place of public examinations for selection purposes.

However, this neatly illustrates the theoretical incompatibility between the use of public examinations of attainment for selection purposes and the use of ability measures to investigate their comparability. The only valid theoretical justification for using ability measures to establish examination comparability, automatically implies that the examinations are inappropriate tools for the selective purpose. On the other hand, if examinations of attainment are appropriate for selective purposes, then the use of ability measures to study comparability is inappropriate. It was noted earlier that the meritocratic philosophy which underpins the use for selection purposes of assessments made by the education system is

more consistent with those selections being made on the basis of attainment than of aptitude. Moreover, in the light of Wood's (1986 and 1991) trenchant criticisms of aptitude testing, this seems just as well. Clearly, if the use of attainment as a selection criterion is accepted, any reference tests which are used to study comparability must be measures of current subject attainment. This brings its own problems which are set out in the next section.

3.4.1.4 The equal-attainment definition

> *Two examinations have comparable standards if, for two groups of candidates with the same distributions of attainment, they produce grades which are identically distributed.*

This definition is perhaps the one which näive observers would give if asked to define comparable grade standards in distributional terms. It is a direct application of the principles of fair meritocratic selection based upon current attainment. It avoids the difficulties caused by the large number of variables which affect attainment by controlling directly for that attainment. However, this definition is essentially circular and assumes a solution for the very problem of assessing different attainments on the same scale which it is intended to solve. It requires comparability first to be established between whatever instrument is used to assess the current attainment of the candidates and each of the examinations being studied before the comparability of the examinations themselves can be investigated. How is this to be done, other than by a further independent assessment of the candidates' attainment? This question makes the infinite regress involved in this definition obvious. The definition was, nonetheless, the basis of some of the Schools Council comparability studies of the 1970s which used attainment tests as reference instruments. However, these studies were discontinued because of the impossibility of constructing such a test which could be independently shown to be equally relevant to the differing assessment domains of the examinations being studied (Forrest and Shoesmith, 1985). Newbould and Massey (1979) gave a convincing demonstration that exactly the same problems arise if the reference instrument is a common element of the examinations themselves.

As noted earlier, some cross-moderation studies have also implicitly adopted this definition. Although they do not involve formal comparisons of grade distributions, some such studies have attempted to define a common criterion variable for use by the examiners when they are judging the quality of scripts from the examinations being studied. Comparable grade

standards are then defined as the award of the same grade, on average, to candidates judged to have the same attainment on the common scale. This is effectively the equal-achievement definition given above. This approach is clearly subject to exactly the same theoretical problems as the use of a reference test. Christie and Forrest (1981) demonstrated the practical consequences of these by re-analysing Houston's (1980) data and concluding (Page 6) that:

> "...no conclusions should be drawn on the basis of an equally weighted composite [of three agreed criterion variables - MJC] for the simple reason that each board differs in the emphasis it accords each criterion..."

3.4.1.5 The similar-schools definition

> *Two examinations have comparable standards if two groups of candidates who attend similar schools receive grades which are identically distributed after studying their respective syllabuses and taking their examinations.*

This definition assumes that, on average, "similar schools" are identical in terms of the school variables relevant to candidate achievement (for example, general effects, effectiveness of subject teaching and entry policy). The two groups of candidates are also assumed to be identical in terms of relevant student variables (prior achievement, motivation) together with the effects upon these of the two syllabuses and examinations. Clearly, much hinges upon how similar the "similar schools" are and the extent to which controlling for schools also controls for relevant student variables. Delta analysis (Quinlan, 1993) is an analytic technique based upon this definition which has recently been used by the examining boards and groups in some of their comparability studies. However, with the data presently routinely collected about the schools and colleges which enter candidates for public examinations, large differences are known to exist within "similar schools". Although this is a problem of current practice, rather than theory, the implicit control of student variables using the school variables as surrogates is an unavoidable limitation of this definition.

3.4.1.6 The catch-all definition

> *Two examinations have comparable standards if two groups of candidates with the same distributions of ability and prior achievement who attend similar schools with identical entry policies, are taught by equally competent teachers and are equally motivated, receive grades which are identically distributed after studying their respective syllabuses and taking their examinations.*

This definition is included here, not because it has been used in any published public examination comparability study, but to raise the question of why it has not been used.

Although it remains prey to the theoretical difficulties of any definition of comparability which depends upon the identity of grade distributions because it, perforce, assumes that the effects upon the student and school variables of the two syllabuses and examinations are identical, this definition is the logical extension of previous statistical work on comparability. The collection of explanatory data about schools and students which the use of this definition would involve has been done in many other studies (for example, Brimer *et al*, 1978; Cresswell and Gubb, 1987), would not be difficult in practical terms and would enable the issue of subgroup comparability to be explored. The only comparability work which seems to have been done along these lines is that begun by Nuttall and Armitage (1984) but similar work on the comparability of public examination standards would be well worthwhile and could make use of recent advances in multi-level modelling techniques (see Goldstein, 1995).

3.4.2 The social value definition - comparable standards defined in terms of value judgements

The foregoing review of statistical approaches to the study of examination grade comparability reveals the existence of several different methodologies, each with its own implicit definition of comparable standards. However, none of these methodologies and definitions is satisfactory within measurement theory and the need for comparability studies to assume that meaning can be given to quantitiative comparisons between qualitatively differing attainments has often caused concern. For example, Johnson and Cohen (1983) say:

> "A situation can be envisaged where this diversity [between assessment domains - MJC] is so great that there would be no consensus among subject specialists that the schemes concerned were equally valid under the same subject title; in such cases the issue of comparability becomes meaningless."

Most authors (for example, Forrest and Shoesmith, 1985) have agreed with Johnson and Cohen that the continuing utility of the public examination system as a provider of information for selection purposes depends upon the possibility of making sufficiently good **approximations** to valid comparisons between attainments in differing domains. A pragmatic case can be made for this point of view when the issue is comparability between different examinations in the same subject but it leaves as meaningless the notion of comparability across different subjects. One of the major consequences of this was the cessation of any

large-scale published British work on comparability between examinations in different subjects from the mid 1970s until Fitz-Gibbon and Vincent's (1994) study which adopted the value-added definition. Nonetheless, the examining boards and groups have continued to claim, if only by implication, that the same grade represents the same standard of attainment in any subject and, in general, selectors have continued to behave accordingly. Recently, the British legal requirement of publication of public examination results as indicators of the success of individual schools has given a new importance to defining comparable standards across subjects because, otherwise, different mixes of subjects taken by different schools will affect the rankings of those schools in the published tables.

In this section, a new definition of comparable standards is proposed which, it is argued, provides an explanation for the long-term success of the public examination system as a provider of information for selection purposes, despite the apparent theoretical impossibility upon which it is based. The new definition also provides a theoretically coherent meaning for the notion of comparability between examinations in different subjects and accomodates the essential requirement, discussed at the start of Section 3.4.1, to consider the relative value of the assessment domains of examinations whose comparability is an issue.

Public examination grades can be likened to a currency with which candidates buy entry into education or employment. Developing this analogy a little further, the intrinsic value of banknotes does not match their face value but commerce functions because it is commonly agreed to accept them at face value. (This agreement is greatly strengthened, but not guaranteed, by the underwriting of a national bank.) Similarly, educational and vocational selection processes can proceed provided that there is common consent that comparability exists between the grades from different public examinations. That is to say, provided that it is accepted that a given grade from one examination represents attainment of equal value to the attainment which earns the same grade in other examinations. From this starting point, it is possible to define comparable standards in social terms involving human judgements of value:

> *Two examinations have comparable standards if candidates for one of them receive the same grades as candidates for the other whose assessed attainments are accorded equivalent value by awarders accepted as competent to make such judgements by all interested certificate users.*

3.4.2.1 Theoretical coherence at the price of subjectivity

There are several important points to note about this definition of comparable standards. First and foremost, it is theoretically coherent because it avoids the formal impossibility inherent in the notion of quantitative equivalences between qualitatively differing attainments which undermine definitions rooted in the theory of educational measurement. This is because the new definition does not require such attainments to be quantitatively equivalent. Instead, by taking seriously the identification of awarding as an evaluative process which was made in Section 3.2, the new definition requires equivalence to relate only to the **value** given to those attainments by the awarders.

The nature of value judgements has long been an area of considerable philosophical debate (see, for example, Pole, 1961; Bambrough, 1979; Best, 1985). However, one issue in that debate is particularly significant in the context of the comparability of examination standards. This is whether value judgements ascribe a property (or properties) to the objects being judged. As far as judgements of educational attainments are concerned, French *et al* (1987) maintain that they do not. Among many others, Ayer (1946), Fogelin (1967) and Billington (1988) have argued the same case in general, albeit from quite different perspectives. Under the social value definition of comparable standards, examination grades, by reporting value judgements, therefore report human responses to the pupils' measured attainment, rather than the attainment itself. From this perspective, there is no **formal** impediment in the way of assigning equal value to qualitatively dissimilar attainments.

However, a price has been paid for this belated philosophical justification of the notion that comparable standards can be defined across differing assessment domains: the need to accept that there is no external and objective reality underpinning the comparability of results from different examinations. The subjectivity which is involved is no more and no less than that which, it has already been argued, is an inherent part of all standard setting methods. As Section 3.3.2 makes clear, such subjectivity need not necessarily lead to capricious or unreliable (in the sense of being difficult to replicate) judgements. However, under the social definition of comparable standards, the legitimacy of using public examination results as the basis for selection depends upon general acceptance that the judgements of examiners are valid and accurate. The next sub-section explores this matter of acceptance in more detail

and addresses, in particular, the question of whether acceptance might be undermined by an explicit recognition of the fundamentally subjective nature of examination standards.

3.4.2.2 User acceptance

It is important to note the *relativistic* aspect of the proposed social value definition of comparable standards. By referring to the *acceptance* of *all interested certificate users*, it implies that other users in other times and places might dissent from the awarders' evaluations of pupils' attainment. Comparable standards as defined here can, therefore, only be established in a particular social context and, for users who do not accept the awarders' competence, will not be achieved. Clearly, the larger the group of examination users who are prepared to accept the awarders' competence to make the required value judgements, the more useful are the examination certificates. The group of users prepared to accept the evaluations of the awarders must include most candidates, parents, teachers and selectors if the examination system is to fulfil its purpose effectively.

This is not to say that, given the opportunity, all users would necessarily make value judgements identical to those of the awarders; indeed, they need not agree with the awarders' evaluations at all, as long as they agree to abide by them. In practice, however, acceptance is only likely to be forthcoming from any particular user on a continuing basis as long as the awarders' evaluations differ only to some small extent from their own judgements, however informal or uninformed the latter may be. Those setting grade standards must therefore either attempt to represent the views of most users or persuade the users that the standards which they set are comparable. The importance which the examining boards have always attached to studying and reporting upon comparability can be seen as an implicit recognition of these requirements. It also demonstrates expertise in the techniques of educational assessent and thereby strengthens the claim of competence to set standards which is implicitly made by the examining boards.

In practice, acceptance of the competence of examining boards to award comparable examination grades appears to be reasonably robust and continues **in general**, despite clear historical instances where some users did not accept their judgements. This was the case,

particularly, for the old CSE and GCE O-level examinations. The examining boards' judgements, based upon their evaluations as supposedly competent judges, that CSE Grade 1 was comparable to an O-level pass were accepted by most teachers and many selectors in education but never fully accepted by selectors in the vocational area. Whatever their own view, some pupils and parents, concerned about subsequent vocational selection processes, therefore behaved as if they, too, did not accept the examining boards' judgement about the comparability of the two systems. The failure of sufficient certificate users to accept the evaluations of the awarders in this case was partly a result of, and partly a reason for, the failure of CSE examinations ever to achieve parity of esteem with O-levels. This contributed to the introduction of the GCSE in 1988 and it may not be unconnected that the GCE boards, which operated the more highly esteemed O-levels, have come to dominate the examining groups which now offer GCSE examinations.

At the time of writing (1996), there is considerable effort being put into achieving parity of esteem between GCE A-level examinations and *General National Vocational Qualifications* (see Dearing, 1996). Whether this is successful will depend upon the reactions of the users of the qualifications but the CSE experience suggests that expedients such as renaming Level 3 GNVQs *Applied A-levels* are unlikely to be effective. It is unknown to what extent the **general** claim of competence by the examining bodies involved will be damaged if users reject the claimed equivalence between these two particular examination systems. However, returning to the earlier analogy with currency, it seems relevant to consider the possibility that examining bodies, whose claim of competence to set accurate standards parallels the underwriting of banks, might be vulnerable, in just the same way as banks, to a catastrophic run on confidence.

Continuing acceptance by users of examination standards in general seems more likely if the procedures used to set the standards are transparent and public knowledge. In the USA, Cizek (1993) has written persuasively that the legal notion of *due process* can be used to underpin standard setting procedures. In Britain, the recent introduction of procedural codes of practice governing public examinations (SCAA 1994 and 1995, the latter having legal force) are a move in this direction. Indeed, the maintenance of user confidence following a critical,

although methodologically extremely dubious, study of GCSE examinations by *Her Majesty's Inspectors of Schools* (HMI, 1992) was the explicit reason for their introduction.

More generally, it seems probable that many users are prepared to accept the standards set by examining boards because they do not believe themselves competent to make the required evaluative judgements for themselves. Indeed, some would argue that the long-term acceptance of public examinations as the principal device for rationing educational resources and vocational opportunity is based upon an unanalysed belief that examination standards have an objective existence. Accordingly, examiners and examination boards are seen as having privileged access to objective standards by virtue of their expertise, rather than being, as is argued here, the people and institutions charged with the responsibility of **constructing** standards on behalf of society as a whole. If belief in objective standards were to be seriously challenged, would general consent for the use of examination results in selection processes be eroded?

No unequivocal answer to this question is forthcoming, but it seems relevant that the continuing public debate about educational standards, although conducted in terms of unshakeable certainties by most of its participants, continually highlights the extent to which different people adopt different views of the standards required of education generally. Even if this is seen by those not directly involved as nothing more than politicians and experts disagreeing with each other, they are visibly disagreeing about values on the basis of opinion, rather than objective facts. Moreover, the presence within the public examination system of regulatory bodies, such as the *School Curriculum and Assessment Authority*, which impose codes of practice, and of appeals procedures which culminate in the quasi-legal proceedings of the *Independent Appeals Authority for School Examinations*, imply acceptance of a need to safeguard candidates from the dangers of errors of judgement and a recognition that, as in legal matters, checks and balances are required to ensure fairness. This is inconsistent with the existence of some widespread belief in the scientific objectivity of examinations as measuring instruments akin to rulers or weighing machines. If the assessment process itself is not seen as objective, but is nonetheless accepted as providing a legitimate basis for selection, it seems unlikely that debate about whether the underlying examination standards

are social constructs or objective realities will have much impact upon the continuing acceptance of public examinations for this purpose.

In any case, it follows from the analysis in this chapter that a truly objective educational assessment system which is consistent with meritocratic principles cannot exist. The repeated failure of strong criterion-referencing provides extensive supporting evidence for this conclusion (and further similar evidence will be added in Chapter 8). Thus, not only are there compelling ethical reasons for making those who are assessed aware of the basis upon which it is done but, from the perspective of examining boards, there is also a pragmatic argument for setting out the impossibility of finding a truly objective alternative. Only if the essentially subjective and culturally determined nature of educational standards is generally understood, can the blandishments of those who offer policy makers simple technical "solutions" to the problems of defining and maintaining examination standards be resisted. Recent work by Fitz-Gibbon and Vincent (1994), which was taken up by Dearing (1996) exemplifies such simple-minded approaches, the methodological poverty and educational dangers of which are discussed in detail by Goldstein and Cresswell (1996).

Finally in this section, it is worth noting that there may be reassurance to be gained from the fact that human judgement is central in the determination of the examination grades which, potentially, have such far reaching consequences in terms of life chances. It opens up the possibility of human concern for individuals within the system and of giving people the benefit of the doubt (which is certainly done, see Chapter 5). Moreover, the subjective nature of human judgement means that examination failure may be easier for the failed to tolerate than it would be if they believed it to represent an objectively true, and therefore incontrovertible, assessment of themselves.

3.4.2.3 Comparability studies rehabilitated

A significant benefit which flows from the theoretical coherence of the proposed social value definition of comparable standards is that it permits comparability to be tested empirically in a theoretically sound way. If comparability is defined in terms of the evaluations of awarders who are accepted by users as competent, there are two empirical questions to ask: are the

awarders' evaluations competent and are they accepted as competent? Taking the second question first, it would be practically difficult but theoretically possible to ask any defined group of certificate users if they accepted the awarders' competence to make the evaluations of attainment implicit in the grades awarded from particular examinations. A less direct test of acceptance would be whether or not the users accepted the examinations concerned as fair. This test exploits the essential reason why comparable standards are important: the need for selection decisions to be meritocratically fair. Indeed, users' views on the fairness of the examination system as a whole could be tested, although they have not been and there would be significant problems of definition to overcome about what represented a sufficient degree of user acceptance to support a general claim about comparable standards throughout the system.

Turning to the issue of the awarders' competence itself, it is a pre-requisite for comparability between specific examinations that the judgements made by the awarders of one examination agree with the judgements of the awarders of a supposedly comparable one. Here the traditional cross-moderation approach to the study of comparability is appropriate and now has a clear theoretical basis although, from the perspective of the social value definition of comparability, cross-moderation studies are concerned with conditions which are necessary but not sufficient to establish comparability. However, under the new definition, comparability between any examinations (including those in different subjects) can, in theory, be usefully studied using cross-moderation methodology. The practical problems of doing so in different subjects would certainly include finding awarders sufficiently knowledgeable to make the required value judgements in more than one subject. The nature of the knowledge required to make awarding judgements was discussed in detail in Section 3.3.2

3.4.2.4 The value of the assessment domain

The final, and possibly most important, point to make about the social value definition of comparable examination standards is that it also deals with the issue of the value of the syllabuses upon which they are based; that is, of the assessment domains. In Section 3.4.1, it was argued that, unless the assumption that syllabuses are of equal value holds, statistical approaches to defining examination standards do so only in a very narrow technical sense

which does not adequately reflect the normal meaning of the term *comparable*. Such a technical definition of comparable standards could imply, for example, that a high typing speed represents a standard of attainment comparable to a post-graduate degree in English. There is not, of course, any **objective** basis for deciding whether or not these attainments, so different in nature, are comparable in standard. However, to assert that they are comparable is not consistent with the value which is currently given to them in British society as measured by the rewards in terms of pay and social status which are given to those holding the jobs for which they act as qualifications (for example, typists and university teachers). Under the social value definition of comparable standards, awarders can, and must, take into account the wider social value of the syllabuses followed by the candidates if they are to make judgements of the value of candidates' attainments which are accepted by users as appropriate for the selective purpose. The necessary and sufficient test of the success of the awarders' attempts to do this is whether or not such acceptance is forthcoming.

3.5 THE APPLICATION OF STANDARDS TO EXAMINATION CANDIDATES' WORK

3.5.1 Two evaluative strategies

When awarders establish a grade boundary for a public examination they are given the task of judging which candidates' work just merits the award of the grade in question. Each candidate's work is judged using relevant evaluative criteria in the way discussed above in Section 3.3.2. However, there are two ways in which the awarders might bring standards to bear upon candidates' work. They might adopt a strategy which can be called *The script as artefact* or they might consider *The script as response*. It is helpful in the analysis of awarders' behaviour in later chapters to distinguish between these two strategies which are therefore considered from a theoretical perspective in this section.

The strategy of *script as artefact* involves the awarders in scrutinising candidates' scripts to determine the sufficient presence of evaluatively relevant qualities. The standards demanded in terms of the evaluative criteria being used relate only to the scripts themselves and do not change from examination to examination. For example, a relevant criterion in Mathematics might be the use of ideas of ratio and proportion and the awarders would look for sufficient

evidence of it which, if found, could become a reason which could be cited in support of their evaluation of the script as a whole. With this approach, there is an implicit assumption that the nature of the assessment tasks does not affect the ease with which the standards can be met; that ideas of ratio and proportion are not easier to apply in some contexts than in others. This is clearly untrue if any tasks at all are considered but nor is it likely to be true of the smaller, more homogeneous group of examination questions which might be set on any particular examination syllabus (see, for example, Pollitt *et al*, 1985).

Nuttall (1987) summarised much of the research evidence relating to the wider context in which the assessment process takes place and Cresswell and Houston (1991) have summarised that relating to the immediate context of assessment tasks. Cresswell and Houston also argue that it is a necessary condition for fairness that context effects should be taken into account during the grading of public examinations.

The strategy of *script as response* accomodates the differing difficulty of assessment tasks by evaluating, not the script *per se*, but the script as a response to a particular set of assessment tasks. Thus, in evaluating candidates' work, the awarders adjust the standard of attainment demanded on each criterion in the light of the difficulty of the particular set of assessment tasks to which the candidates responded. The question thus arises of which strategy is theoretically more desirable.

3.5.2 The implicit interpretative domain score

In Section 3.2.4 it was established that the purpose of the awarding process is to make evaluations which are independent of any particular set of tasks so that selectors can make fair comparisons between candidates who have been assessed by different examinations. According to the DIE model of assessment, independence from a particular examination can only be achieved by evaluating an interpretative report of the candidates' performances in terms of the wider universes of knowledge, skills and understanding defined by the syllabuses. However, conventional public examination procedures do not involve the explicit formation of such an interpretative domain score (see Figure 3.8) so, in effect, the awarders are required

indirectly to evaluate the implications of the candidates' work in terms of the syllabuses by directly scrutinising the work itself.

To achieve this, awarders must form their value judgements in a way which takes into account the particular set of assessment tasks used on the particular occasion and how these relate to the syllabus as a whole. Thus, awarders must not judge pupils' work as if it were a free-standing artefact, but rather as a response to the tasks which provoked it. The *script as response* evaluative strategy is therefore required. This conclusion will be of considerable importance in Chapters 5 and 6 when the procedures and results of awarding meetings are analysed.

3.5.3 Maintaining or defining standards?

In normal public examining practice, an awarding meeting is held for every examination on a particular syllabus (Chapter 1). It is not, however, always made explicit by the examining boards whether the function of these meetings is the successive application of a particular set of independent evaluative standards or simply the application of the grading standards used in the immediately preceding year to the current examination. These two approaches will be called the *definition* and *maintenance* of standards, respectively. Of course, in theory the same result should be produced by either definition or maintenance, assuming that the evaluative process works correctly every time.

3.5.3.1 Maintenance of standards

However, the distinction does have considerable significance. Despite all that has been said in this chapter about awarding as an evaluative process and about definitions of comparability, if the focus is **only** upon the maintenance of standards between successive examinations on an unchanging syllabus, some statistical approaches to comparable awarding become more viable. In particular, the same-candidates definition of comparability (Section 3.4.1.3) offers a theoretically coherent approach. All the assumptions of that approach are reasonable if the same group of candidates, having reached the end of their course, sit two examinations on the same syllabus. Thus, it would be theoretically possible to adopt the same-candidates

definition for the maintenance of comparable standards between successive examinations on an unchanging syllabus.

To do this in practice, it would be necessary to administer last year's examination to a representative sample of candidates, at the time when they sat the current year's. There are obvious practical and methodological matters which would need attention, such as randomising the order of the administration of the two examinations and keeping the previous year's examination unknown to the sample candidates, but none of these is so difficult that it could not be dealt with to a sufficient degree to ensure the credibility of the approach. It would also be essential to investigate the stability of the results of the approach for well-defined subgroups of the candidates (see Section 3.4.1) and this highlights a major difficulty which would need to be overcome: the representativeness of the sample.

Public examination candidates are under considerable pressure when they take their examinations and, characteristically, they take several in a short period of time. This period is therefore one of intensive activity for the candidates, either taking an examination or preparing for the next one. Thus, it could be argued that being in the sample of candidates which took two examinations would put candidates at a disadvantage in their other, subsequent examinations by depriving them of preparation time and, perhaps, increasing their anxiety. Against this, it might be possible to award the sample candidates the better of their two grades, giving them two chances to show their attainment as compensation for any adverse effects upon their other examination results. However, it is clear that, for the reasons just given, the sample used would have to be self-selected. This would raise considerable questions about its representativeness.

An alternative approach would be to administer the two examinations to the sample at a time other than during the operational examination period. This would be practically possible by adopting a rolling program in which the examination taken in Year x was administered with that intended for use in Year x+2 to a sample of candidates due to take their operational examination in Year x+1. However, a problem of sample representativeness would remain with this arrangement. The self-selection effects might be smaller but there would be the

additional problem that the candidates would not be motivated in the same way as they are during the operational examination period and would not be fully prepared for the examinations. It would have to be assumed that any such effects affected both examinations equally, which might not be the case since they would not necessarily cover identical parts of the syllabus.

Despite these difficulties, however, the possibility of maintaining public examination standards in this way would be worth exploring so as to establish the scale of the anticipated problems. Such an *equating* approach is in widespread use elsewhere in the world, notably in the USA where it has spawned an extensive literature (see Holland and Rubin, 1982).

As an alternative to the same-candidates definition, the similar-schools definition of comparability (Section 3.4.1.6) is worth considering for the maintenance of standards between successive examinations on an unchanged syllabus. Indeed, since many schools enter candidates for the same examination in successive years, a *same-schools* definition is possible. With this approach, comparability between two examinations set successively upon the same syllabus would be defined as follows:

> *Two successive examinations are comparable if two groups of candidates who attend the same schools receive grades which are identically distributed after studying their respective syllabuses and taking their examinations.*

The implicit assumptions behind the use of this definition to maintain standards between successive examinations on an unchanging syllabus are that the schools do not change during the period between the two examinations in terms of the various school variables discussed in Section 3.4.1 and that the use of the same schools also controls effectively for the various student variables discussed in that section. It would be possible to research the extent to which these assumptions hold in general. If it was found that they held to an extent which was judged sufficient, then the same-schools definition of comparability would be practically much easier for the examining boards to use than the same-candidates definition. Indeed, the boards already implicitly use the similar-schools definition in their awarding procedures when they compare the grade distributions of the current examination with those from the previous year. However, the degree of emphasis which is given to this approach

varies between the boards and none of them use it **in place of** awarders' judgements (see Chapter 1).

One final consideration about the use of exclusively statistical approaches to the maintenance of standards is worth mentioning. This is the transparency of the awarding process. In general, to use statistical methods is to adopt procedures which many people find difficult to understand and of which some people are suspicious. To say that a group of experts (the awarders) judge the quality of candidates' work to award the grades appears comprehensible and reassuring to a lay audience of pupils, parents, teachers and other interested parties. That it, in fact, appeals to a process of judgement, which is barely understood at all (hence the present study) is obscured by the everyday familiarity of the process of making judgements. However, as noted in Chapter 2, appearance, as well as reality, is an important consideration in choosing procedures for public examining.

3.5.3.2 Definition of standards

It was argued in the preceding section that statistical alternatives to awarders' value judgements might be viable for maintaining standards between successive examinations on an unchanged syllabus. However, this leaves the problem of setting grade standards on the first examination on a new syllabus or after a revision to a syllabus. Here, the arguments in Section 3.4 apply and there is no theoretically coherent alternative to the use of value judgements made by appropriately qualified awarders.

The qualifications needed for the task are twofold. First, Section 3.3.2, implies that the awarders must share tacit standards, based upon guild knowledge, with all those making awards in other syllabuses in the same subject and also share more general tacit standards and guild knowledge with those making awards on examinations in other subjects which report results in terms of the same grade scale. Second, the tacit standards adopted by the awarders must reflect the views of the wider group of examination users (see Section 3.4.2) sufficiently to be accepted by them as comparable with those applied to other examinations reporting on the same scale of grades.

This leaves open the issue of where the tacit standards come from in the first place. They are perhaps best understood as a dynamic norm established within the teaching profession considered as an identifiable group within society (see Brown, 1988, for a discussion of norms within social groups). The norm is dynamic because it clearly changes over time as the curriculum, in its widest sense, develops. It is rooted in teachers' professional experience (of both their pupils' attainments and the way in which these are rewarded in examinations), discussion with their colleagues and contact with educational thinking in general. The dynamic norm which underpins public examination standards does not, therefore, represent an objective yardstick with which changes in the performance of candidates over a long period of time can be measured but it has proved sufficiently stable to provide a basis for selective processes in the educational and vocational worlds for many years.

3.6 CONCLUDING REMARKS

In this chapter it has been argued that public examination awarding is neither a process of norm-referencing nor of criterion-referencing. It is best understood as an evaluation of pupils' attainments. This conception is consistent with the selective function of the examinations which is their primary purpose (see Chapter 2). It also enables a definition of grade comparability to be constructed which, for the first time, provides a philosophically coherent basis for testing the implicit claim made by public examining boards that the same grade represents the same standard of attainment in all assessment domains, no matter how they differ. In Chapters 5, 6 and 9, the extent to which awarding procedures correspond, in practice, to this theoretical model is examined.

CHAPTER 4

AGGREGATION AND AWARDING PROCEDURES

"What would life be without arithmetic, but a scene of horrors?"
- Revd, Sydney Smith

4.1 AGGREGATION

Most public examinations consist of a number of components such as written papers, coursework, multiple choice tests and so on. Performance on these components must be aggregated to produce the overall subject grade. Methods of aggregation can be grouped into two classes (Cresswell, 1988): those which permit compensation to operate freely between components so that candidates can compensate for weakness in one area by strength in another and methods which limit the operation of such compensation by making the award of a particular grade dependent upon achieving some specified level in one or more individual components.

Conventionally, the process of aggregation in public examinations is seen as unproblematic. Numerical marks are assigned to pupils' responses to each question and these are then added together to form a total score for the component. Similarly, the aggregation of the marks from different components within an examination is done by adding together the two component totals, sometimes after multiplying by a scaling factor to adjust the weight which each component exerts in the aggregate. Considerable research has been done on such conventional aggregation, particularly with respect to the issue of the weight which each component exerts within the examination total (for example, see Adams and Murphy, 1982; Cresswell, 1987a and Delap, 1994).

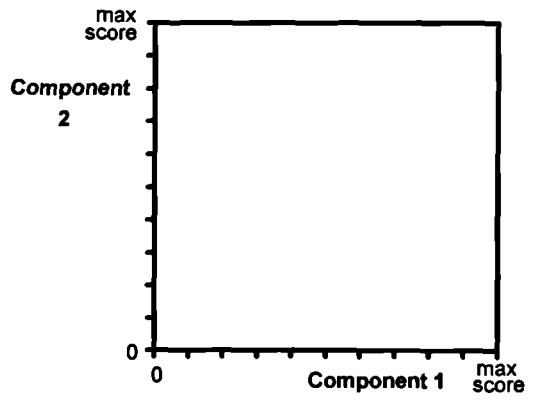
An alternative proposal for aggregation is based upon the use of decision theory within public examination awarding (French *et al,* 1987). It involves the derivation of possibly non-linear scaling functions which are applied to numerical examination component marks before they are added in a conventional way to form an examination total. The scaling functions are derived from examiners' judgements of the relative merits of sample pupils' work.

Further new approaches to aggregation have been proposed in connection with the development of notions of strong criterion referencing. The concern of much of this work has been to aggregate in such a way as to preserve descriptive information from the lower level in the assessment hierarchy so that the more general higher level description allows low level information to be inferred. Wiliam (1995a and 1995b) discusses approaches of this sort within a general theoretical framework; Cresswell (1987b and 1994) discusses them in two particular practical contexts.

4.1.1 The score space

A useful conceptual tool for considering various approaches to the aggregation of scores from assessments is the space created by arranging orthogonally the measurement scales of the component assessments which are being aggregated. The score space for a public examination with two components is shown in Figure 4.1. For clarity, the discussion in this section will be set in the context of this simple example but it generalises directly to aggregation involving more components.

Figure 4.1
Basic score space for an examination with two components



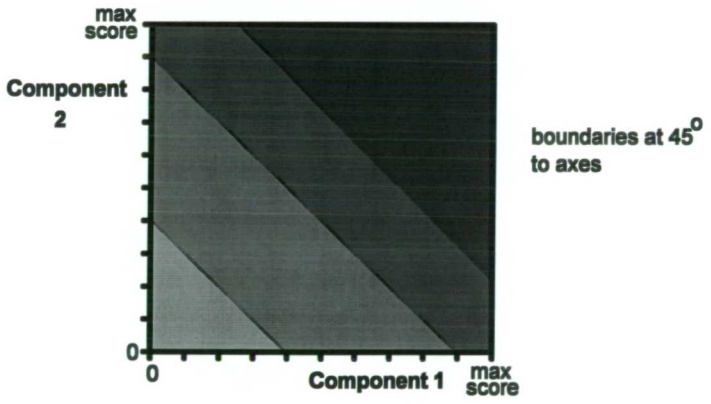
Different aggregation methods can be illustrated on the score space by the shape of regions which indicate increasing attainment in terms of the aggregate. In the diagrams which follow, four such illustrative regions are defined but this number is arbitrary. Indeed, just as it is usually assumed that the attainment measured on each component of the score space is

continuously distributed, so the aggregate attainment is best considered to be continuously distributed for most purposes. However, if the divisions on the score space axes are thought of as ordered categories, the score space can also be used to describe procedures for combining component grades or the categorical data which come from strongly criterion-referenced assessments. This approach will be used in Section 4.2 below.

4.1.2 Conventional aggregation

Conventional aggregation by the unweighted addition of raw component marks is shown in Figure 4.2.

Figure 4.2
Score space showing conventional aggregation of equally weighted components
(Darker shading indicates higher aggregate attainment)



In Figure 4.2, the boundaries between regions of increasing aggregate attainment are straight lines at angles of 45° to the axes. As a result, marks from the two components are completely interchangeable within the aggregate. Pupils' total scores, alone, determine their aggregate attainment with a good performance on one component able to compensate fully for a poor one on another component. The aggregation function illustrated is $A = c_1 + c_2$, where A is the aggregate attainment and c_1 and c_2 are the raw scores on Components 1 and 2, respectively.

Figure 4.3 is the same as Figure 4.2, except that the raw component scores are no longer equally weighted. The angle of 30° between the boundaries and the Component 1 axis

implies that Component 1 is given half the weight of Component 2 in determining pupils' aggregate attainment (see Cresswell, 1987a).

Figure 4.3
Score space showing conventional aggregation of components weighted 1:2
(Darker shading indicates higher aggregate attainment)

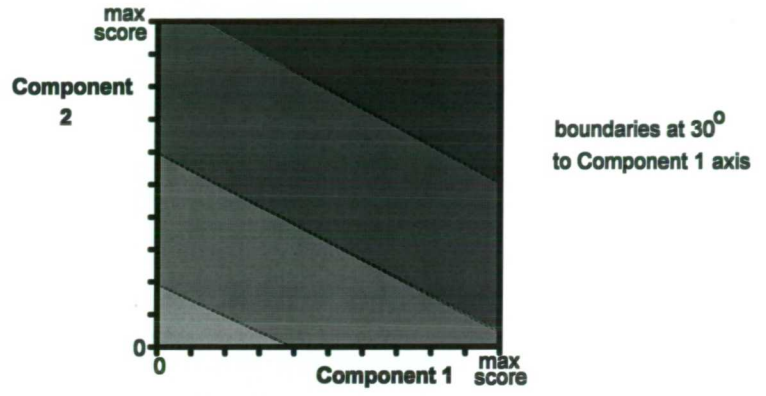


Figure 4.3 illustrates a particular case ($A = c_1 + 2c_2$) of the general aggregation approach used in public examining. In general, the aggregation function used in public examinations is:

$$A = \sum_{i=1}^{n} \omega_i \cdot c_i$$

Where ω_i is the scaling factor for Component i and there are n components. Note that the scaling factor ω_i is not necessarily equal to the intended weight (w_i) of component i because, in the general case, all components are not necessarily scored out of the same maximum raw mark (M_i). In general,

$$\omega_i = \frac{w_i}{M_i} \cdot C$$

where C is an arbitrary constant.

Although the boundaries shown in Figure 4.2 and Figure 4.3 are parallel, it is possible to justify aggregation schemes in which they are not. Christie and Forrest (1981) suggested that it might be reasonable in some circumstances to argue that the relative weights given to the different components of an examination should be allowed to vary across the aggregate grades. For example, it might be argued that the knowledge and skills assessed in a theory component were relatively more important *vis-a-vis* a practical component for the award of the

highest aggregate grade than for the lower grades. Christie and Forrest (1981) reported data, based upon a re-analysis of a study by Houston (1980), which show how in one O-level English Literature examination different aspects of attainment were given different emphasis at different grades by awarders from different boards. However, in normal practice, the physical components of an examination do not necessarily assess the different facets of attainment to which it might be desired to attach differing weights at different levels. This practical problem is a major impediment to the adoption of Christie and Forrest's general approach.

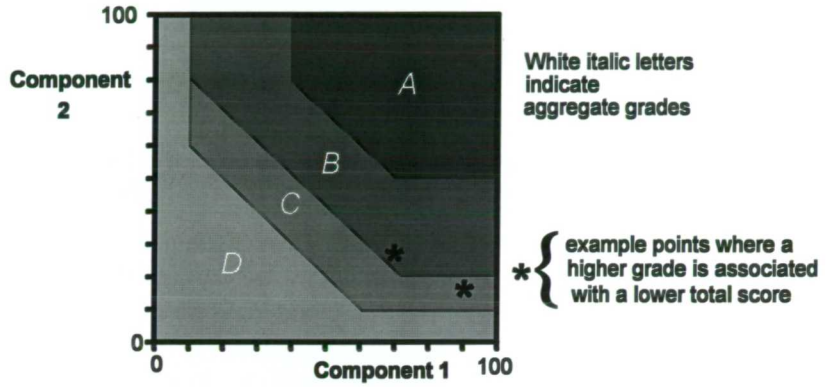
It is important to note that the interpretation of the slopes of the aggregate attainment boundaries in the score space in terms of component weights assumes that the component scores form an equal interval scale (Cresswell, 1987a). This assumption is also fundamental to many of the mark transformations routinely carried out by examining boards as French *et al* (1987) note. French *et al* challenge the equal mark interval assumption but it is difficult to see what empirical test of it could be devised. Since no scale of attainment can be constructed independently of a particular assessment instrument, there is no independent criterion available to judge the equality of the intervals on any particular mark scale. Only appeals to theoretical considerations about learning in the subject being assessed can possibly shed light on this question but public examinations are designed with such considerations in mind and, in at least one case, those devising question papers and marking schemes are explicitly asked to use such considerations to make every mark "equal in value" (SEG, 1988). On this basis, it can be argued that unscaled component marks are on equal interval scales **by definition**.

4.1.3 Hurdles

One of the ways in which compensation within conventional public examination aggregation has sometimes been modified is by the use of grade hurdles. These are scores on particular components which must be obtained before a particular aggregate grade is awarded. Figure 4.4 illustrates a number of the characteristic features of grade hurdles. The hurdles are not necessarily equally spaced either within or between the components and two grades (B and C) share the same hurdle (10 marks) on Component 1. (Note also that, in this case, the

differently shaded areas correspond exactly to the aggregate grades for which the hurdles apply since hurdles are defined in terms of qualification for particular grades.)

Figure 4.4
Score space showing the operation of grade hurdles with two otherwise equally weighted components
(Darker shading indicates higher aggregate attainment)



In recent years, the use of grade hurdles in British public examinations has diminished. There are two main reasons usually given for this. First, practical hurdles affect few candidates. In essence, grade hurdles operate to penalise pupils whose performance on the examination's components is very disparate. However, since there is generally a moderate positive correlation between examination components, most pupils' marks fall in the regions of the score space where normal aggregation operates. (This can be seen from Figure 4.4 by imagining a scatterplot of positively correlated pupils' scores superimposed upon it.) Second, grade hurdles make the grading process less transparent (see Chapter 2) since some pupils have higher aggregate grades than others, but lower total scores. This effect is caused because the hurdles interfere with compensation between disparate performances on different components. It is illustrated by the two points marked with asterisks on Figure 4.4. Concern about the public acceptability of this property of grade hurdles (despite it being their very purpose!) has tended to lead, in practice, to them being placed at very low mark levels, further reducing the number of pupils affected.

4.1.4 The Decision Theoretic approach

It is clear from the foregoing that grade hurdles penalise pupils whose performance is uneven across the components of the examination. The aggregation approach developed by French *et al* (1987) permits the use of aggregation functions which achieve similar ends without the discontinuities evident in Figure 4.4 and follows on from Christie and Forrest's (1981) work on the definition of aggregate grades in terms of varyingly weighted facets of attainment. French *et al* propose the use of decision theory techniques to derive aggregation functions on the basis of a number of arbitrary technical assumptions and judgements made by examiners about the relative merits of particular sample pupils' scripts. Figure 4.5 and Figure 4.6 show the score spaces for continuous aggregation functions which reward even and uneven performances respectively.

Figure 4.5
Score space showing a continuous aggregation function which rewards even performance

(eg. partially hyperbolic function: $A = c_1 + c_2 + \frac{c_1 c_2}{k}$)

(Darker shading indicates higher aggregate attainment)

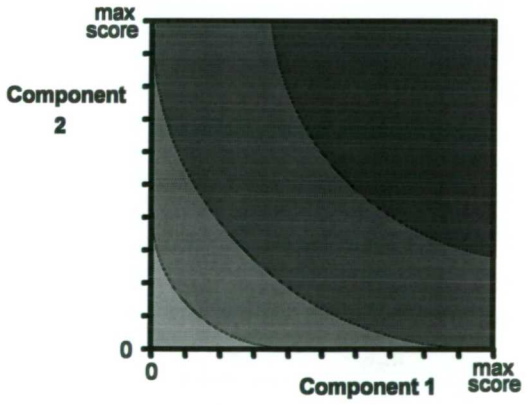
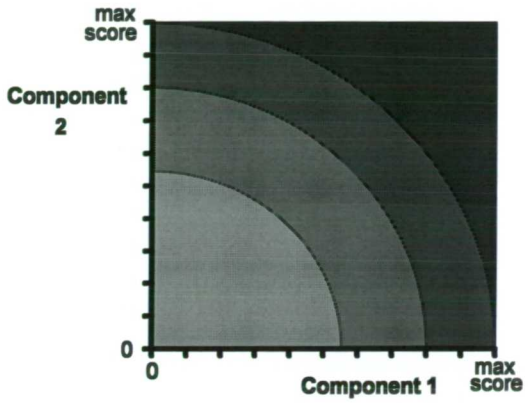


Figure 4.6
Score space showing a continuous aggregation function which rewards uneven performance

(eg partially circular function: $A = c_1 + c_2 + \frac{c_1^2 + c_2^2}{k}$)

(Darker shading indicates higher aggregate attainment)



For illustrative purposes, Figure 4.5 and Figure 4.6 greatly exaggerate the size of effects which French *et al* advocated (or believed to be plausible on the basis of examiners' judgements). Despite its flexibility, their approach has not been adopted by the examining boards because of the lack of transparency of both the aggregation functions derived and the method of deriving them. Aggregation functions of the types illustrated in Figure 4.5 and Figure 4.6 are therefore of primarily theoretical interest as examples of the effect of modifying the way in which disparate component performances can compensate for each other.

4.2 COMBINING COMPONENT JUDGEMENTS

As described in Chapter 1, in conventional awarding (see SCAA 1994 and 1995), evaluative judgements are made of candidates' work on each component of the examination separately, rather than on the examination as a whole. There are several reasons normally given for this practice. First, it is held that awarders are better able to evaluate candidates' work on individual components than on the examination as a whole. One justification for this view is that different components normally focus on different aspects of the subject being examined so that component judgements involve the use of a smaller range of evaluative criteria than would judgements of performance on the whole examination and therefore reduce the

complexity of the awarders' task. Similarly, awarders generally report greater difficulty in arriving at a judgement of a candidate's work if, within it, there are both very good and very poor aspects which need to be traded off against each other. By judging each component separately, the extent of this difficulty is reduced.

The second reason normally given for making awarding judgements separately on each component is that some components are part of more than one examination. Although there are theoretical arguments in favour of judging performance on a component differently in the differing contexts of the examinations to which it contributes, the need for transparency (Section 2.9.1) is normally taken to require the same judgement to be made of a given performance on a component, regardless of the examination in which it is embedded. Evaluating performances on each component separately ensures that this is done. Finally, the third reason given for making awarding judgements separately on each component is a practical one. Modern examinations often involve coursework or practical components which are too physically bulky to be collected for all candidates. It is not, therefore, always possible to assemble the complete examination work of an individual candidate for evaluation in the awarding meeting. By considering the components separately, only independent samples of work are required for each component, simplifying their collection considerably.

The awarders' judgements thus produce a set of boundary marks which partition each **component** mark scale into regions corresponding to grades. The question then arises of how these judgements should be combined to determine aggregate grades for candidates in terms of the **subject** as a whole. One obvious approach is to award grades on each component and then combine these grades. Alternatively, a method is needed of deriving grade boundary marks in terms of the aggregate mark scale from the boundary marks on the component mark scales. In this section, these various approaches are considered.

4.2.1 Aggregating component grades

As indicated in Section 4.1.1, if the continuous mark scales shown on the axes of the score space are replaced by component grade scales, then awarding processes involving the

combination of component grades, rather than component marks, can be modelled. Grade combination, rather than mark aggregation, has been advocated most strongly in three main contexts.

4.2.1.1 Component Grade Profiles

The first context in which component grade aggregation offers some advantages over component mark aggregation, is when component grades are to be reported alongside aggregate grades. This is done to provide a more specific account of pupils' attainment than is possible with a single aggregate grade. Most of the GCE Boards and GCSE Groups provide uncertificated component grade profiles for this purpose. One difficulty which arises, however, is that, when only the component and aggregate grades (ie. not the marks) are reported, there appear to be anomalies between pupils' component grade profiles and their aggregate grades. Table 4.1, which is reproduced from Cresswell (1988), illustrates how such apparent anomalies can arise.

Table 4.1
How apparent anomalies arise in a profile grading system
(taken from Cresswell, 1988)

Suppose an examination has two components with total marks of 40 and 60:

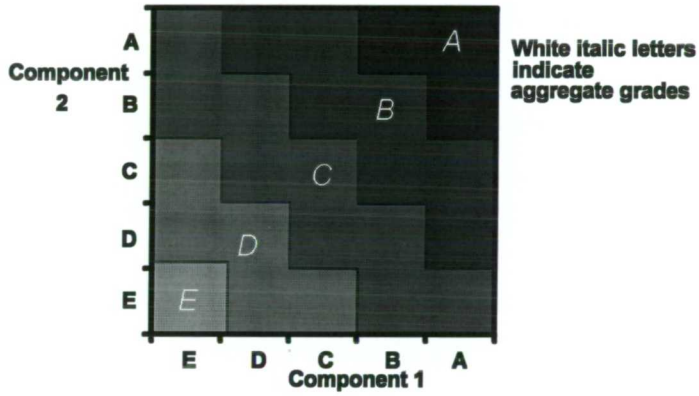
Practical	Max. mark = 40	Grade A/B boundary = 35
Theory	Max. mark = 60	Grade A/B boundary = 50
Total	Max. mark = 100	Grade A/B boundary = 85

It is then possible for three candidates to have marks and grades as follows:

		Practical	Theory	Total
Candidate 1	marks	38	49	87
	grades	A	B	A
Candidate 2	marks	35	47	82
	grades	A	B	B
Candidate 3	marks	34	49	83
	grades	B	B	B

In Table 4.1, Candidates 1 and 2 have the same component grade profiles but different aggregate grades; Candidates 2 and 3 have different profiles but the same aggregate grade; indeed, as a set, Candidate 2's grades are better than Candidate 3's even though Candidate 3 has a higher aggregate score. Clearly, no such apparent anomalies would occur if the aggregate grade were to be determined by combining the component grades, rather than the component marks. In this case, a score space like that shown in Figure 4.7 would typically be used.

Figure 4.7
Typical score space for grade aggregation
(Darker shading indicates higher aggregate attainment)



Since the boundaries in Figure 4.7 are, on average, at 45° to the axes, the components are given equal weight by the aggregation rules shown. Figures corresponding to Figures 4.3 and 4.4 can easily be constructed to illustrate the cases of unequally weighted components and hurdles under grade aggregation.

However, there is a price to pay, in terms of the reliability of the aggregate grades, for avoiding apparent anomalies by grade aggregation. There is a strong relationship between the number of grades used to report the component attainments and the reliability of the aggregate grades. The only published analysis of this relationship is Cresswell (1988) where it is argued that, for grading systems to be acceptable, there must be no candidates who are more likely to be wrongly graded than correctly graded. This principle always holds for grades based upon mark aggregation but it is shown in Cresswell (1988) that it is generally true under grade

aggregation only if the number of component grades is higher than the number of aggregate grades.

Unfortunately, practical grade aggregation systems rarely meet this requirement since the same grade scale is usually used to report both component and aggregate attainment. For this reason, the GCE and GCSE examining boards have generally adopted a policy of attempting to explain apparent anomalies between profile and aggregate grades, rather than adopting grade aggregation. This policy is driven by the view that the reliability of the aggregate grade should be paramount, given its selective use (Chapter 2).

4.2.1.2 Modular schemes

The second context in which grade aggregation has been advocated is for assessment schemes in which component grades are earned by pupils over a period and eventually "cashed in" for an aggregate grade. Modular examinations where each module is assessed upon completion are the most common case in point, exemplified by the Open University Student Handbook which indicates the final degree class which will be awarded for every possible set of contributing course outcomes.

In addition to avoiding apparent anomalies like those illustrated in Table 4.1, grade aggregation has two additional advantages for modular schemes. The first of these is that pupils can work out the aggregate grade which their present module grades would produce and decide, on this basis, whether to cash them in or to re-take or take further modules in the hope of obtaining a better aggregate grade later. The second advantage is that the grade scale used to report the modules has the effect, by definition, of putting assessments from different modules onto the same scale. If this is effective, aggregation for pupils who have taken different sets of modules is greatly simplified. These advantages have evidently been judged by the Open University to accrue and to be worth the reliability costs of grade aggregation. On the other hand, concern about reliability has led the GCE and GCSE boards generally to adopt mark aggregation systems for modular examinations, despite the consequential need to equate marks from modules before they are aggregated. Chapter 7

reports some empirical work on the results of grade aggregation in a modular A-level examination.

4.2.1.3 Strong criterion-referencing

The third context in which it is sometimes argued (for example by Ward, 1980) that grade aggregation offers particular advantages is where it is desired to be able to make inferences from the aggregate grade to attainment in terms of the components. Strongly criterion-referenced assessments are a case in point, exemplified by the British driving test where there are seven separately assessed objectives, for **all** of which a satisfactory performance must be demonstrated before an overall pass is awarded (Wiliam, 1995a). However, the nature of the inferences about component performances which can be made depends crucially on the method used to aggregate the component grades.

Consider, for example, the aggregation system shown in Figure 4.8. This is a *conjunctive* grade aggregation scheme in which a given aggregate grade is only awarded to pupils who achieve at least that grade on Component 1 **AND** Component 2. (The first use of the terms *conjunctive* and *disjunctive* in this context was by Christie, 1982.) It is often argued that a conjunctive scheme is capable of "preserving a high degree of criterion-referencing" (as Wiliam, 1995b, puts it, for example) by which is meant, presumably, that strong inferences about component performance can be drawn because each aggregate grade carries with it a guarantee that at least that grade was achieved in every component. However, the strength of the inferences which can actually be drawn is evident from Figure 4.8. Aggregate Grade A is, indeed, completely unambiguous in terms of its implications about pupils' component grades but the other aggregate grades become increasingly ambiguous as their distance from Grade A increases. At the other extreme from Grade A, Aggregate Grade E can be awarded to pupils with profiles as disparate as A, E and E, A.

Figure 4.8
Score space for a conjunctive grade aggregation system, claimed to permit strong criterion-referenced inferences (Darker shading indicates higher aggregate attainment)

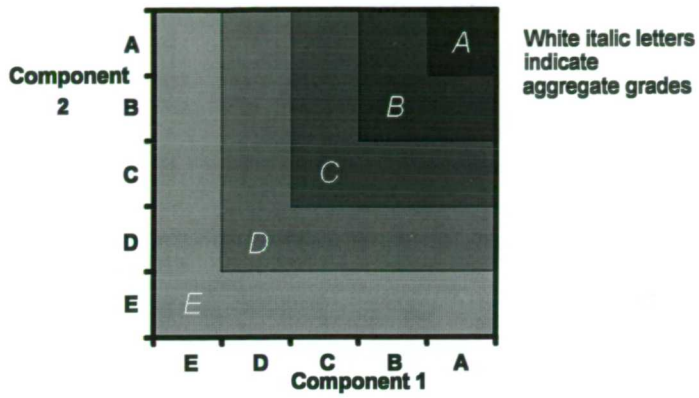
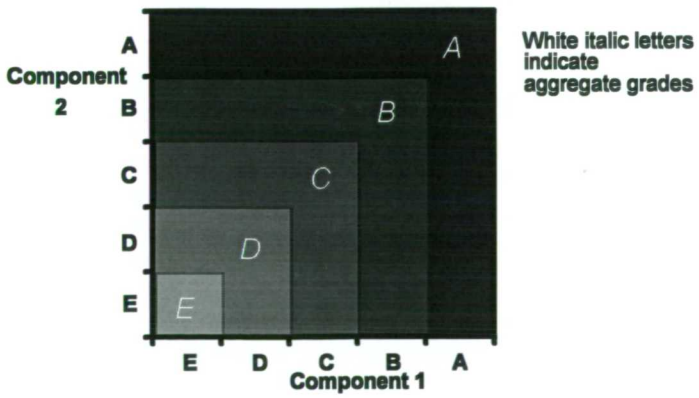


Figure 4.9 illustrates a *disjunctive* grading scheme in which the aggregate grade awarded is the grade achieved by the pupil on the component where he or she does best. If Figures 4.8 and 4.9 are compared with Figures 4.5 and 4.6, respectively, it can be seen that conjunctive schemes reward pupils whose performances are even across the components and disjunctive schemes reward those whose component performances are uneven. It is also worth noting that, since there is no detraction from a good performance on one component by a poor performance on another, disjunctive schemes might entice pupils to concentrate their efforts solely upon the content of one component. Wiliam (1995a and 1995b) discusses several aggregation schemes from this consequential perspective.

Figure 4.9
Score space for a disjunctive grade aggregation system
(Darker shading indicates higher aggregate attainment)



Interestingly, although disjunctive schemes are rarely advocated by the proponents of strong criterion-referencing, the extent to which specific inferences about component attainments can be drawn from the aggregate grades which they produce is no different from that for a conjunctive scheme. As a comparison of Figures 4.8 and 4.9 makes clear, only the location of the ambiguity in the aggregate grades is different.

Why, then, are conjunctive schemes sometimes advocated by virtue of the component inferences which they permit? The answer seems to lie in erroneous argument from analogies such as the driving test and a failure to appreciate that reporting in terms of a scale of several grades is fundamentally different from reporting simply in terms of the two categories *pass* and *fail*. In the latter case, an aggregate pass, awarded conjunctively, implies passes in all components and the fact that an aggregate fail does not enable which components were passed or failed to be inferred is of no interest to the users of certificates, precisely because, as failure, it is not certificated. In a graded examination, on the other hand, failure just to obtain, say, Grade B means that Grade C is certificated and inferences about component attainments will still be of interest to the users of the certificates.

In Cresswell (1988) an analysis is given of the two classes of compensatory and non-compensatory aggregation schemes. It is concluded that grade combination systems which limit the operation of compensation, such as those illustrated in Figures 4.8 and 4.9, carry with them a substantially increased likelihood that some candidates will receive an incorrect aggregate grade. In Cresswell (1988) it is observed that limiting compensation might be desirable in some contexts:

> "If an overall pass qualifies candidates to practice surgery for example, it may be thought desirable [by inhibiting compensation] to stack the odds against awarding incorrect passes even at the expense of increasing the likelihood of incorrect failures."

However, in a general educational context, if it is desirable to report specific information about candidates' strengths and weaknesses within a subject, this is most effectively done in the form of a component grade profile. An accompanying aggregate grade is then most useful if it records each candidate's mean attainment since it is superfluous if it is simply a coding

corresponding to several different profiles. This requires an aggregation method which permits compensation to operate freely between the various examination components. A conclusion similar to this was reached by Thyne (1974) using different, but related, arguments.

4.2.2 Combining component boundaries

For reasons outlined in Section 4.2.1.1, compensatory aggregation of components produces the most reliable result if it is done by aggregating component marks, rather than component grades, and this is the method used in all non-modular GCSE and GCE examinations. It is therefore necessary to devise a method for combining component grade boundaries so as to produce grade boundaries which partition the aggregate mark scale into regions corresponding to aggregate grades. In terms of the score space made by the scaled component marks, and assuming conventional aggregation, a method is required for determining the position of the 45° lines which correspond to aggregate grade boundaries. In current practice (SCAA, 1994 and 1995) two approaches are used.

4.2.2.1 Aggregating component boundaries

The obvious approach is simply to aggregate the component boundary marks in the same way as the candidates' marks are aggregated, as follows:

$$B = \sum_{i=1}^{n} \omega_i \cdot b_i \qquad \text{Model 1}$$

where B is the aggregate boundary for the grade in question and b_i is the corresponding raw score boundary for Component i.

Prior to the introduction of GCSE examinations, Model 1, which is sometimes called the *addition method,* was the approach generally used to combine component boundaries. Under it, candidates must meet the standard set for a particular grade on every component (or the equivalent of this, allowing for compensation) to be awarded the grade on the aggregate. Thus, the implicit question which awarders setting standards on each component in turn must address is: *what standard of work, when aggregated with a comparable standard of work on all the other components, will result in an aggregate standard which merits a Grade x?*

4.2.2.2 Equating aggregate and component score scales

Equating aggregate and component score scales is the alternative approach in use. Under it, the component boundaries are not seen as setting a series of distinct standards which must be met as a whole but as separate indications of the quality of work associated with the grade in question. The equivalents, in terms of aggregate scores, of these indications are then averaged to determine the standard required for the aggregate grade. The implicit question which awarders must address for each component is therefore: *what standard of work, on this component alone, merits a Grade x?*

A version of this approach, based upon equipercentile equating, has increasingly been used since 1988. This version, known as the *percentile method*, defines the aggregate boundary for Grade x as the mark which corresponds, on the aggregate score distribution, to the weighted mean of the cumulative percentages of candidates awarded Grade x on the components. The weights used to form the mean are the intended weights of the components. Thus, the aggregate boundary is given by

$$B = F^{-1}\left[\frac{\sum_{i}^{n} w_i \cdot f_i(b_i)}{\sum_{i}^{n} w_i}\right] \qquad \text{Model 2}$$

where f_i is the cumulative distribution function for Component i, F is the cumulative distribution function for the aggregate and w_i is the intended weight of Component i.

4.2.2.3 Technical discussion of the two approaches

Methods for the determination of grade boundaries on conventional aggregate score scales from component grade boundaries have been little discussed in the literature. The only technical accounts of them which have been identified are in Good and Cresswell's reports (1988a and 1988b), recommendations of good practice by the GCE boards' Standing Research Advisory Committee (SRAC, 1990) and Quadling's (1992) article in the context of Mathematics examinations (although his points are largely general ones). All these authors considered only Models 1 and 2, above, since these are the only ones ever used in practice in British public examinations. They considered the bunching effects due to regression to the mean which occur in the aggregate score because the component marks are not perfectly

correlated. The usual consequence of these effects is that, above the mean, the use of the addition method means that more candidates exceed a given grade boundary on each component than on the aggregate; below the mean, more candidates exceed a given grade boundary on the aggregate than on each component.

Responses to these phenomena differ. Quadling (1992) concluded by recommending the addition method, quoting Thomas Carlyle's remark that *the world is a republic of mediocrities* to justify the acceptance of the aggregate consequences of regression to the mean. On the other hand, SRAC (1990) recommended the use of the percentile method **when it gives a lower aggregate boundary than the addition method** in order to compensate for what they considered the tendency of awarders to give inadequate consideration to the accumulating effect of their individual component decisions (this was dubbed the *tunnel vision* effect). SRAC therefore interpreted regression to the mean effects on the aggregate scores as a measure of this tunnel vision effect. However, SRAC did not recommend the use of the percentile method when it gives a **higher** aggregate boundary than the addition method for reasons of transparency. They argued that it would be difficult to justify some of the results of doing this; for example, a candidate who had exceeded the pass mark set on every component of an examination might, using the percentile method alone, fail the examination as a whole. Current practice for both GCSE and GCE examinations (SCAA 1994 and 1995) broadly follows the SRAC recommendations.

Good and Cresswell (1988b), writing before the percentile method had been used operationally in British public examinations, said that the addition method is appropriate if the components all measure the same trait (except for random error) and that the percentile method is appropriate if the components measure different traits. The heuristic argument which they gave in support of this assertion is that, in the latter case, the aggregate grade boundaries should be closer to the mean than the weighted sum of the component boundaries because the regression to the mean effects in the aggregate score scale are caused by the partial cancellation of component-specific attainment factors within the aggregate score. Good and Cresswell acknowledged that, in practice, imperfect correlation between two examination components cannot be identified solely with either measurement error or trait

differences. They went on to recommend that the choice of boundary combination method should depend upon the judgement made about the predominance of one or other source of imperfect correlation in each individual examination. They also recommended that, once determined, the aggregate boundaries should then be "adjusted to take account of the source of error excluded from the model that is used [to derive them]" using examiners' qualitative judgements. The details of this adjustment process were not worked out.

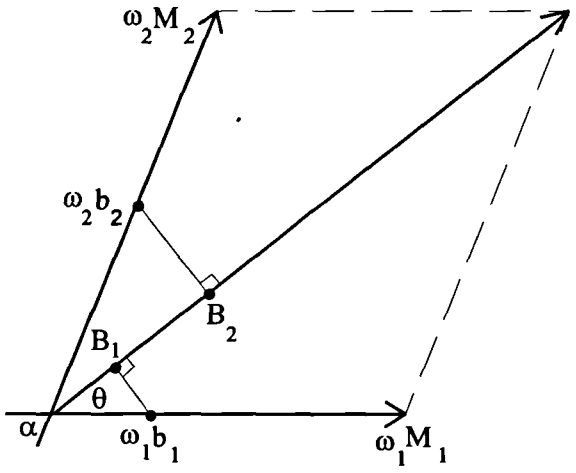
In common with SRAC (1990), Good and Cresswell (1988b) also did not address the question of whether the percentile method gives the right adjustment for the regression effects which they used to justify it. Good and Cresswell were working solely with the two methods known, at that time, to be under consideration for the new GCSE examinations in 1988 and had an essentially pragmatic agenda. Two years later, in a similarly pragmatic vein, SRAC were considering applying those same procedures at A-level. A more thorough theoretical analysis of methods for combining component boundaries is thus long overdue and is provided in this section.

As SRAC (1990) realised, the essence of the mattter lies in the behaviour of the awarders. If the component boundaries are decided in the light of their aggregate implications, then the addition method is appropriate. However, if the awarders judge performance on each component independently, using the addition method creates significant difficulties. These are best illustrated by reflecting on the consequences for the aggregate standard of adding another component to an existing examination. In practice, the new component will not be perfectly correlated with the existing ones. As a result, if the awarders make boundary decisions on the new component which are comparable with, but do not take account of the aggregate effects of, the boundary decisions on all the components, regression to the mean effects will result in fewer candidates being able to meet each of the new aggregate boundaries above the aggregate mean and more candidates being able to meet each of those below the aggregate mean. In the limit, by adding more and more components and grading them all comparably but independently, the situation could be reached where all candidates were awarded the one grade which straddled the aggregate mean! Thus, if the addition method is used to combine component boundaries set independently, without reference to

their collective consequences, the grade standards set for the examination as a whole depend, in part, upon the number of components in the examination. This is clearly undesirable and a different method for combining component grade boundaries is needed if these are set independently of each other, whether by design or due to the effects of *tunnel vision*.

One way of exploring the adjustments which might be made to aggregated boundaries for regression to the mean effects between the components is to model the component and aggregate scores as vectors. For clarity, the following argument refers in detail to a two component examination, but general results are given for any number of components. Figure 4.10 shows a theoretical two component examination in vector terms.

Figure 4.10
Theoretical vector representation of a two component examination



The imperfect correlation between the components is represented by the angle α between them. For the moment, we will take the vectors to represent the true score variables, rather than the observed score variables, so that Figure 4.10 represents Good and Cresswell's second case of imperfect correlation due to differences in the traits assessed by the components. The aggregate true score is represented by the resultant vector which makes an angle θ with the Component 1 vector. The lengths of the component vectors are the maximum marks for the components **after any scaling has been carried** out: $\omega_1 M_1$ and

$\omega_2 M_2$. The points $\omega_1 b_1$ and $\omega_2 b_2$ are the scaled component boundary marks for the grade in question, with $\vec{B}_1$ and $\vec{B}_2$ the component vectors of these along the aggregate vector.

The resultant of the two component boundary vectors, given by:

$$\vec{R} = \overrightarrow{\omega_1 \cdot b_1} + \overrightarrow{\omega_2 \cdot b_2}$$

will lie along the aggregate vector if and only if:

$$\frac{b_1}{b_2} = \frac{M_1}{M_2}$$

If this is not true, then the component boundaries set by the awarders do not reflect the intended weights of the components. There are a number of possible responses to this state of affairs. Cresswell (1987a) pointed out that the component weights implied by component grade boundary decisions can be used as the basis for an alternative measure of the achieved weights of the components which is independent of the candidates' performances. On the other hand, Christie and Forrest (1981) suggested that the awarders might legitimately want the components to have different weights at different grade boundaries (this was discussed in Section 4.1.2). For the purposes of the present analysis, however, we shall take the intended aggregate variable to be the one which should be used to grade the candidates, since this is the syllabus designers' intention, so that the length of the component of R which lies along the aggregate vector defines the aggregate boundary corresponding to the composite of b_1 and b_2. From the geometry of Figure 4.10, this component is easily shown to be given by:

$$\vec{B} = \vec{B}_1 + \vec{B}_2$$

Which implies:

$$|\vec{B}| = \omega_1 \cdot b_1 \cdot cos\,\theta + \omega_2 \cdot b_2 \cdot cos\,(\alpha - \theta)$$

Again from the geometry of Figure 4.10, and remembering that the cosine of the angle between two vectors is equal to their correlation, this gives:

$$|\vec{B}| = \frac{\omega_1^2 \cdot M_1 \cdot b_1 + \omega_2^2 \cdot M_2 \cdot b_2 + r_{T12} \cdot \omega_1 \cdot \omega_2 \cdot (M_1 \cdot b_2 + M_2 \cdot b_1)}{\sqrt{\omega_1^2 \cdot M_1^2 + \omega_2^2 \cdot M_2^2 + 2 \cdot r_{T12} \cdot \omega_1 \cdot \omega_2 \cdot M_1 \cdot M_2}}$$

where r_{T12} is the true score correlation between Components 1 and 2.

The addition method for combining component boundaries (Model 1) can be seen to be a special case of this general aggregation method which arises when the component true scores are perfectly correlated so that $\alpha = 0$; then $|\vec{B}_1| = \omega_1 b_1$ and $|\vec{B}_2| = \omega_2 b_2$. This is consistent with Good and Cresswell's (1988b) suggestion that the addition of the scaled component boundaries is appropriate when imperfect correlation between the components is due only to measurement error.

However, the above analysis assumes that the maximum marks on the components accurately represent their relative weights. In practice, this is unlikely to be true for a variety of reasons discussed in detail in Cresswell (1987a). The above analysis also implicitly treats the component marks as forming ratio scales because of the intersection of the component vectors at zero. Such an assumption is unnecessary and, almost certainly, incorrect in practice. An alternative formulation, based more closely upon the practice of examining is shown in Figure 4.11 where s_{ri} is the standard deviation of the raw (unscaled) marks on Component i and the component marks have been centred upon their raw (unscaled) means, m_{ri}.

Defining the aggregate boundary once again as the sum of the vector components, along the resultant aggregate vector, of the component boundaries and using the statistics of the components and aggregate, the aggregate boundary for n components is:

$$B = \sum_{i}^{n} \omega_i \cdot (b_i - m_{ri}) \cdot r_{Tia} + m_a \qquad \text{(since, eg., } \cos\theta = r_{Tia}\text{)}$$

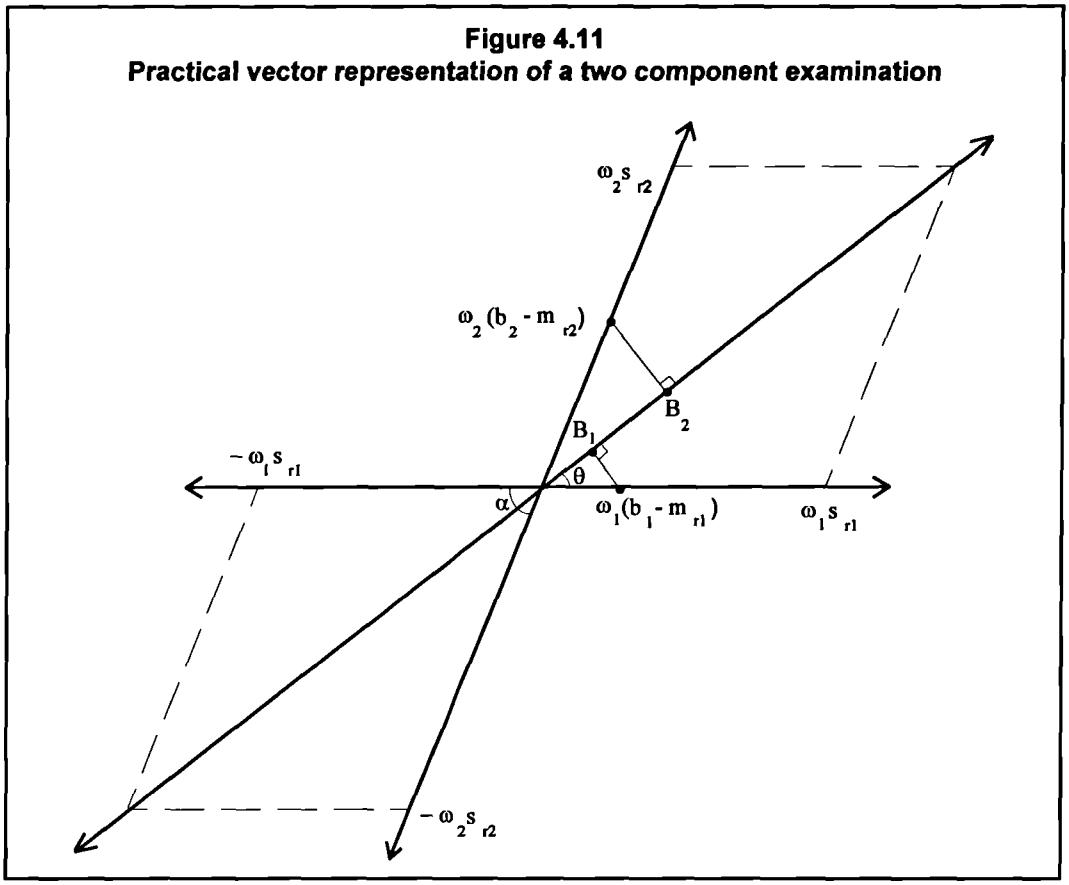
where r_{Tia} is the true score correlation between Component i and the aggregate and m_a is the mean of the aggregate scores.

Expressing r_{Tia} in terms of component statistics gives:

$$B = \frac{\sum_{i}^{n} \left[\omega_i \cdot (b_i - m_{ri}) \cdot (s_i \cdot r_{ii} + \sum_{k \neq i}^{n} r_{ik} \cdot s_k) \cdot \sqrt{\prod_{k \neq i}^{n} r_{kk}} \right]}{\sqrt{\sum_{i}^{n} [(s_i^2 \cdot r_{ii} + \sum_{k \neq i}^{n} r_{ik} \cdot s_i \cdot s_k)]} \cdot \sqrt{\prod_{i}^{n} r_{ii}}} + m_a \qquad \text{- Model 1a}$$

where s_i is the standard deviation of the scaled scores on Component i, r_{ii} is the reliability of Component i and r_{ik} is the observed correlation between Components i and k.

Figure 4.11
Practical vector representation of a two component examination



Model 1a for aggregating component boundary decisions is the combined model which Good and Cresswell (1988b) argued was required but did not derive. Model 1a allows for only those regression to the mean effects which are due to the cancellation of component-specific true variance within the aggregate score. In practice, however, examination component reliabilities are not routinely determined so the routine use of Model 1a would require assumptions about them to be made.

The following model corresponds exactly to Model 1a but relates to the observed, rather than true, scores and therefore makes allowance for imperfect component inter-correlations whatever their source.

$$B = \frac{\sum_{i}^{n} \omega_i \cdot (b_i - m_{ri}) \cdot (s_i + \sum_{k \neq i}^{n} r_{ik} \cdot s_k)}{\sqrt{\sum_{i}^{n} (s_i^2 + \sum_{k \neq i}^{n} r_{ik} \cdot s_i \cdot s_k)}} + m_a \qquad \text{Model 1b}$$

All the data required to use this model are routinely available, so it could be used in practice to avoid the problematic effects of regression to the mean effects which occur if component grade boundaries are set without consideration of their aggregate effect.

An interesting feature of Model 1b is that it defines the aggregate boundary as the weighted mean of the z scores corresponding to the component boundaries, using weights based upon the conventional component-with-aggregate covariance measure (Adams and Murphy, 1982). Thus, Model 1b can be seen as fitting within the general equating approach to the definition of aggregate boundaries described in Section 4.2.2.2. However, Model 1b uses a conventional linear equating function (Holland and Rubin, 1982; Good and Cresswell, 1988b), rather than the equipercentile approach of Model 2.

In fact, Model 2 (the percentile method) is the version of the equating approach which is used in practice. One immediate question about Model 2 is raised by the preceding analysis: why is the mean percentage formed by weighting the component percentages in accordance with the **intended** weights of the components? The conventional justification for weighting the component percentages **at all** is a pragmatic one: that intended weights reflect the relative importance attached within the syllabus to the attainments assessed in each component and that the percentage of candidates getting Grade x for the subject as a whole should most closely reflect the percentages getting Grade x in the more important components. For example, it is not thought appropriate for a high percentage of candidates getting Grade x in a relatively unimportant component to lead to a percentage of candidates getting Grade x for the subject as a whole which is much higher than those obtained on the more important components. The use of achieved, rather than intended, weights in Model 2 does not seem to have been previously considered. There are transparency advantages in using the intended weights but, as will shortly emerge, potential disadvantages as well.

A more technical justification of the use of a weighted mean percentage was suggested by Good and Cresswell (1988b) who showed that, with an unweighted mean, the percentile method does not always give grade boundaries closer to the mean aggregate score than the addition method. That it should do so, was seen as a pre-requisite by Good and Cresswell

because they viewed the percentile method as a way of adjusting the aggregate boundaries for regression to the mean effects. The present analysis suggests that this is an appropriate view only when the weights used are the component-with-aggregate covariances, rather than the intended weights. Indeed, it will shortly be shown that the intended weights do not necessarily give aggregate boundaries which meet Good and Cresswell's pre-requisite.

A mathematical criticism which can be offered of Model 2 is that the process of forming a weighted mean of two percentages is invalid since such percentages do not form an equal interval scale (Adams, 1993). This problem is easily avoided by using Model 2a which averages the aggregate marks equivalent to the component boundaries, rather than the percentages of candidates to which the component boundaries correspond.

$$B = \frac{\sum_{i}^{n}\left[w_i \cdot F^{-1}\{f_i(b_i)\}\right]}{\sum_{i}^{n} w_i} \qquad \text{Model 2a}$$

A variant of Model 2a, equivalent to Model 1b, is obtained by replacing the intended weights with the component-with-aggregate covariance weights, as follows:

$$B = \frac{\sum_{i}^{n}\left[(s_i^2 + \sum_{k\neq i}^{n} r_{ik} \cdot s_i \cdot s_k) \cdot F^{-1}\{f_i(b_i)\}\right]}{s_a^2} \qquad \text{Model 2b}$$

where s_a is the standard deviation of the aggregate scores.

In order to explore the similarities and differences between these models for combining component boundary judgements, their results have been evaluated for simulated two-component examinations in which the unscaled component marks are assumed to be normally distributed on a scale of 0 to 100 with means of 50 and standard deviations of 12. The effects on the results of each model of varying correlation between the two components were explored for two different sets of intended component weights: 1:1 (scaling factors $\omega_2 = \omega_1$) and 1:2 (scaling factors $\omega_2 = \omega_1 \cdot 2$). The results of this work are shown in Figures 4.11 and 4.12.

Figure 4.11
Aggregate boundaries using combination Models 1, 1b, 2, 2a and 2b plotted against inter-component correlation for a two component examination; component boundaries at 65 and 70 out of 100, component unscaled scores distributed N(50, 12), equal intended weights (scaling factors $\omega_2 = \omega_1$)

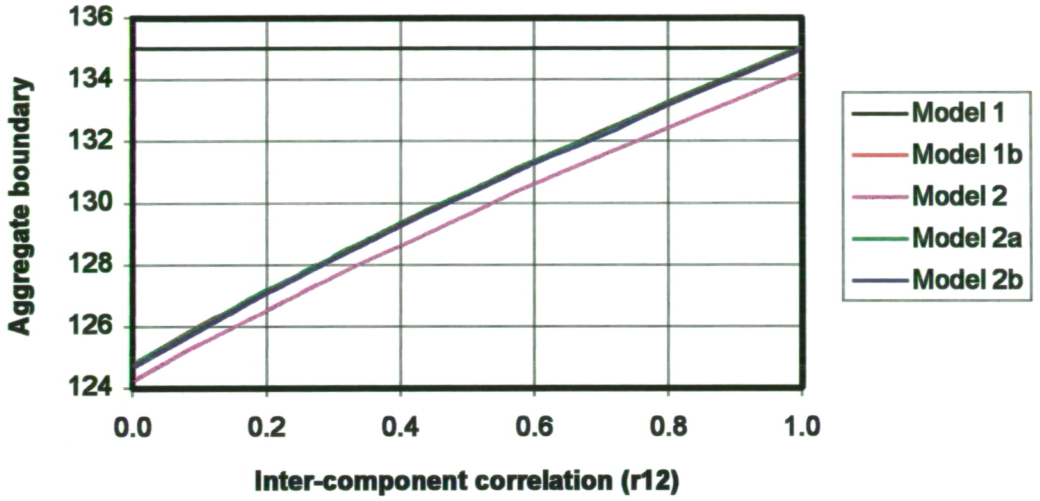


Figure 4.12
Aggregate boundaries using combination Models 1, 1b, 2, 2a and 2b plotted against inter-component correlation for a two component examination; component boundaries at 65 and 70 out of 100, component unscaled scores distributed N(50, 12), intended weights 1:2 (scaling factors $\omega_2 = \omega_1 \cdot 2$)

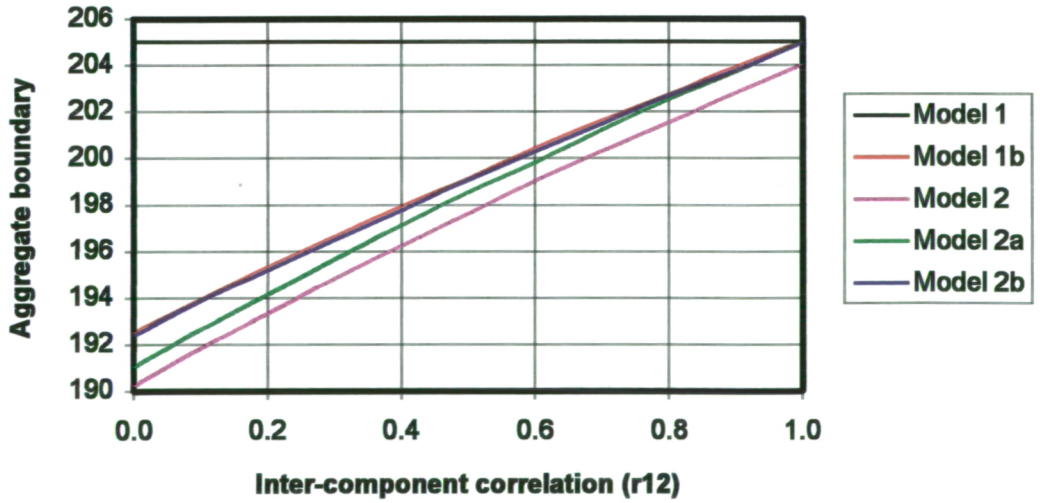


Figure 4.13
Aggregate boundaries using combination Models 1, 1b, 2, 2a and 2b plotted against inter-component correlation for a two component examination; component boundaries at 65 and 70 out of 100, component unscaled scores distributed N(50, 12) and N(50, 20), equal intended weights (scaling factors $\omega_2 = \omega_1$)

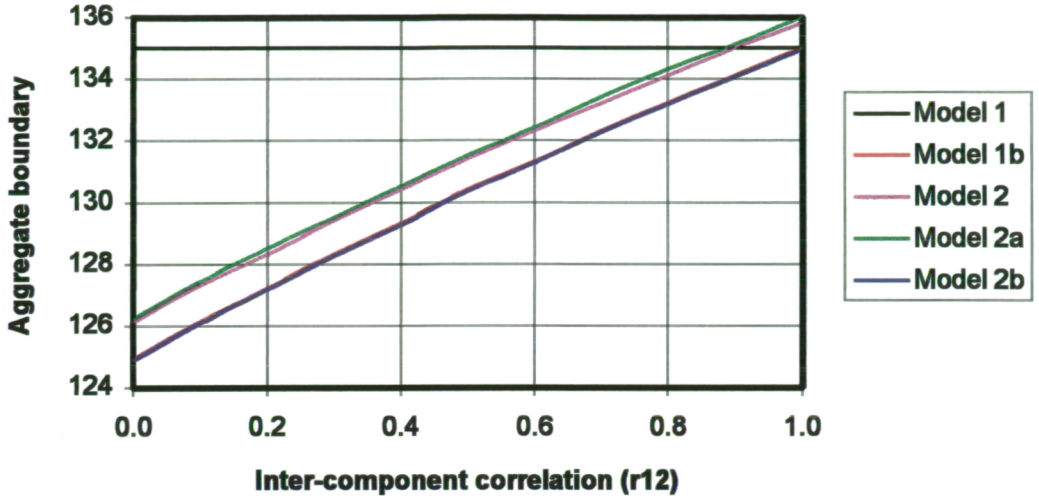
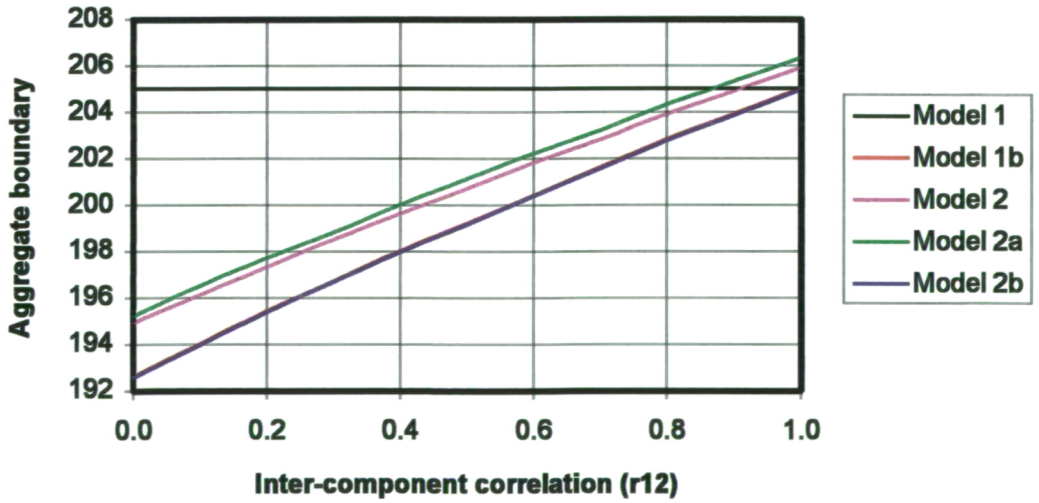


Figure 4.14
Aggregate boundaries using combination Models 1, 1b, 2, 2a and 2b plotted against inter-component correlation for a two component examination; component boundaries at 65 and 70 out of 100, component unscaled scores distributed N(50, 12) and N(50, 20), intended weights 1:2 (scaling factors $\omega_2 = \omega_1 \cdot 2$)



A number of points can be seen from Figures 4.11 and 4.12. First, for these examinations, Models 1, 1b, 2a and 2b all give the same aggregate boundaries when the correlation between the components is 1. This is to be expected for Models 1 and 1b since, when $r_{ik} = 1$ $\forall$ i, k,

$$\frac{\sum_{i}^{n} \omega_i \cdot (b_i - m_{ri}) \cdot (s_i + \sum_{k \neq i}^{n} r_{ik} \cdot s_k)}{\sqrt{\sum_{i}^{n} (s_i^2 + \sum_{k \neq i}^{n} r_{ik} \cdot s_i \cdot s_k)}} + m_a = \sum_{i=1}^{n} \omega_i \cdot b_i$$

In Figure 4.11, Models 1b, 2a and 2b all give the same aggregate boundaries for any component inter-correlation (the curves have been slightly separated in the figure for visibility's sake). Models 1b and 2b give the same aggregate boundaries in Figures 4.11 and 4.12 because, with normally distributed marks, linear equating (Model 1b) and equipercentile equating (Model 2b) are necessarily equivalent. Models 2a and 2b are equivalent to each other for examinations in which the achieved and intended weights are identical and this is true for the examinations in Figure 4.11 because the intended weights are 1:1 and the component marks have identical scaled standard deviations.

Models 2a and 2b are not, however, equivalent across the range of correlation in Figure 4.12 because, with unequal scaled component standard deviations, the component-with-aggregate covariance measure of achieved weight moves further away from the ratio of the standard deviations as the inter-component correlation approaches zero. Clearly, Model 2 should also give the same result as Model 1 when the inter-component correlation is perfect if the rationale for Model 2 is to make allowance for regression effects (which must, under these conditions, be zero). Unfortunately, it does not.

In general, of course, the condition of identical component raw score standard deviations which is illustrated in Figures 4.11 and 4.12 is unlikely to be met in practice. However, it is important to realise that the way in which the analyses in this section are framed means that differences between component raw score standard deviations which **precisely** reflect the ratio of the component raw mark maxima are also not an issue. Such differences are accommodated within the scaling factors ω_i and are the reason why it cannot be assumed that

the $\omega_i = w_i$, the intended weights of the components. As a result, systematic differences of this type do not affect the picture presented in Figures 4.11 and 4.12.

The effect of **incidentally** unequal unscaled component standard deviations upon the aggregate boundaries given by the different models is still an issue and is shown for two simulated examinations in Figures 4.13 and 4.14. Note, in particular, that when the ratio of the unscaled standard deviations does not exactly reflect the ratio of the raw mark maxima, Models 2 and 2a can give aggregate boundaries further from the mean aggregate score than Model 1 (in these examples, when the inter-component correlations exceed about 0.9). This is clearly an unsatisfactory feature of these models if their *raison detre* is to correct for regression to the mean effects by moving the aggregate boundaries nearer to the mean aggregate score. This feature of the models concerned arises because of their use of the intended weights as weighting factors and does not occur for Models 1b and 2b which use the component-with-aggregate covariance weights.

4.2.2.4 Applying the models

To explore their behaviour for real examinations, Models 1, 1b, 2, 2a, and 2b have been applied to the 13 A-level examinations with more than 5000 candidates offered by one board in Summer 1993. Model 1a could not be tried since no reliability data were available for the examinations in question. The results are contained in Table 4.2; statistical details and component boundaries for the examinations in question are to be found in Appendix 4.1.

Table 4.2
Five different models for combining component grade judgements applied to 13 A-level examinations, each with over 5000 candidates.
(see Appendix 4.1 for further statistical details)

Subject (aggegate mean)	Grade	Model 1	1b	2	2a	2b
Biology (126.6)	A	176	172	169	170	172
	B	151	150	144	145	149
	E	99	103	101	100	102
Business Studies (121.1)	A	179	172	171	171	171
	B	153	150	149	149	149
	E	103	105	105	105	105

Table 4.2
continued

Subject (aggegate mean)	Grade	Model 1	1b	2	2a	2b
Communication Studies (243.3)	A	330	310	305	307	307
	B	282	273	269	271	271
	E	200	211	212	211	212
Economics (87.6)	A	138	131	129	131	130
	B	114	111	110	110	110
	E	72	74	73	73	73
English I (156.1)	A	220	207	206	206	206
	B	191	184	183	184	184
	E	123	130	129	129	129
English III (209.8)	A	272	261	260	260	260
	B	242	236	234	234	235
	E	145	156	158	156	157
French (234.1)	A	318	302	298	299	299
	B	269	262	260	260	260
	E	175	186	186	185	185
Physics A* (287.8)	A	383	376	371	372	374
	B	345	341	335	336	339
	E	243	248	244	243	245
Physics B* (281.6)	A	386	378	371	372	375
	B	347	344	335	338	342
	E	247	252	248	247	249
Psychology (242.1)	A	353	337	333	334	335
	B	306	296	295	295	296
	E	191	200	202	199	201
Pure and Applied Mathematics (92.4)	A	152	149	148	149	148
	B	126	124	124	125	124
	E	71	71	72	72	72
Sociology I (85.8)	A	126	122	121	122	121
	B	111	109	109	109	109
	E	77	78	78	78	78
Sociology II (100.9)	A	138	131	131	131	131
	B	117	114	114	114	113
	E	86	89	88	88	88
Theatre Studies (133.4)	A	191	178	171	174	175
	B	165	158	154	155	156
	E	108	115	116	114	115

* Physics A and B are two variants of a single examination in which two versions of the Paper 3 practical are set.

4.2.2.5 The model of choice

It is apparent from Table 4.2 that there is little difference between the results of the models which allow for regression effects (1b, 2, 2a and 2b) although Model 1b consistently produces slightly more severe results than the others. On this evidence, there is no empirical basis for choosing between the models.

In practice, the decision as to which model to use to combine component boundary judgements must balance considerations of technical merit, practicality and transparency. The technically best model cannot be decided without reference to the implicit questions which awarders setting component boundaries consider. In Sections 4.2.2.1 and 4.2.2.2, these questions were set out as follows: *what standard of work, when aggregated with a comparable standard of work on all the other components, will result in an aggregate standard which merits a Grade x?* (aggregation models) and *what standard of work, on this component alone, merits a Grade x?* (equating models). In general, these questions should not produce the same answers but examining board procedures (see Chapter 1) do not explicitly ask awarders to address either one or other of them. In these circumstances, since the choice of a model for combining component boundaries depends upon the implicit question which awarders are trying to answer, it is necessary to try to find out what that question is before the general approach for combining component boundary judgements can be chosen. Some evidence relevant to this issue is reported in Chapters 5 and 6.

A further consideration which needs to be borne in mind is that any grade boundary combination method which uses the statistics of the candidates' marks introduces dependencies between the candidates' grades. Using such methods, the aggregate grades of some individuals could be different if they took the same examination among a different group of candidates, even if the same component grade boundaries were used. This is clearly undesirable. Such effects may be small when large numbers of candidates are involved but might be considerable for some small entry examinations. Only Model 1 avoids this problem.

If, however, the implicit question with which awarders concern themselves implies that allowance for regression effects should be made, then the use of models which introduce dependencies between candidates' results is unavoidable. (The alternative of establishing

regression allowances from a pre-test is theoretically possible and should be considered. However, its desirability would have to be evaluated in the light of the technical difficulties which would arise related to the pre-test samples - see Section 3.5.3.1 - and its unprecedented practical implications for the examining boards' other procedures and timetables.) Of the models which make allowance for regression, Model 2b is to be preferred on theoretical grounds. It avoids the mathematically questionable averaging of percentages of Model 2, the assumption of a linear relationship between component scores which is made by Model 1b and the use of intended weights, regardless of the characteristics of the mark scales in practice, which is a feature of Model 2a. Note, however, that for examinations with relatively few candidates, Model 1b might be preferable in practice because it depends only upon summary statistics so, for small samples, is likely to be more robust than an approach involving equipercentile equating.

In terms of practicality, Model 1 is the easiest to use. Model 1b, because, it uses summary statistics and does not require reference to the details of mark distributions, is also relatively straightforward. Models 2, 2a and 2b, are likely to be more complex in practice. However, Model 2 is already in routine use, so the practical difficulties of any of the models seem likely to be manageable, given adequate computing support. More significant, perhaps, is the transparency of the procedures discussed in this section. Model 1 is probably the easiest for a lay audience to understand and accept. Transparency problems with Model 2 will be reported in Chapter 5 and the other models are certainly not easier than Model 2 to explain to a non-technical audience. A final recommendation on the preferred model for combining component boundary judgements will be deferred until Chapter 9, after data on the nature of the judgements made by awarders have been reported.

4.3 THE NATURE OF THE COMPONENT

In this chapter, it has been assumed that the components being aggregated, and for which grade boundary judgements must be combined, are conventional examination components, each assessing material relevant to the award of all the grades and defined either by the assessment techniques used (eg. objective test, essay paper, coursework and so on) or by

the subdomain of knowledge, skills and understanding which they assess. There are two types of examination component which differ from this general paradigm: *differentiated* components and those used in recent National Curriculum assessments.

4.3.1 Differentiated components

The first type of unconventional component occurs in GCSE examinations where, in some subjects, candidates can choose from among different sets of components which lead to different ranges of grades. Such differentiated components operate, within the different versions of the examination which they define, exactly like conventional components. The above analysis therefore applies to them without modification. Clearly, serious additional issues of comparability arise when the same grade is awarded from two different versions of the same examination and these were the focus of Good and Cresswell's (1988a and 1988b) study.

4.3.2 National Curriculum assessment components

The second unconventional type of component is that used, in various forms, in recent British National Curriculum assessments. For example, in the 1993 Key Stage 3 assessments (Ruddock,*et al*, 1993), each question answered by the pupils was identified with a particular National Curriculum level and the sets of questions associated with each level were treated as separate components. Aggregation within these components was conventional and pupils were awarded the level associated with a component if their mark exceeded a pre-determined mastery score set on it. With the intention of facilitating strongly criterion-referenced interpretations (see Chapter 3), the aggregate awarding procedure then involved a disjunctive rule, the aggregate level awarded to a pupil being the highest component level awarded to that pupil regardless of whether lower levels had been awarded. Cresswell (1994) evaluates this approach to aggregation and awarding from a theoretical point of view. Chapter 8, below, reports an empirical investigation of it.

4.4 CONCLUDING REMARKS ON AGGREGATION AND AWARDING

In Section 4.1.1 of this chapter, a theoretical framework - the score space - was offered for understanding the aggregation and awarding methods which are in common use in British public examinations. The framework provided by the score space has also proved useful in the analysis of less conventional aggregation and awarding approaches. In Section 4.3, a detailed study was made of methods of combining component boundary judgements so as to position aggregate grade boundaries appropriately within the conventional score space. Some improved models for doing this have been developed and applied to both simulated and real examinations. It has been argued that the appropriateness of the models depends upon the details of the judgemental process by which awarders arrive at the component grade boundaries. To illuminate this and other matters, an observational study of those processes is reported in the following two chapters. Subsequently, Chapters 7 and 8 report empirical studies of grade aggregation and aggregation within strongly criterion-referenced awarding procedures.

CHAPTER 5

A QUALITATIVE ANALYSIS OF CONVENTIONAL AWARDING: THE OBSERVATIONAL WORK PART 1

"The bearings of this observation lays in the application of it."
- *Dombey and Son*, Charles Dickens

5.1 PURPOSE OF THE OBSERVATIONAL WORK

In the preceding chapters a theoretical perspective on the process of awarding public examination grades has been developed which places value judgements made by certain judges, known as awarders, at its heart. The aim of the present research as a whole, the critical evaluation of awarding practices, requires a clearer description and deeper understanding of these judgemental processes than has hitherto been generally available. This chapter, and the following one, describe an observational study of conventional examination awarding meetings which was mounted with the aim of providing the necessary description and understanding of the way in which human judgement is used in the awarding process. In particular, it was hoped to identify:

1 the evidence which is used by awarders as a basis for their judgements;

2 the ways in which awarders use that evidence;

In Chapter 9, the evidence used by the awarders, and their ways of using it, are evaluated from the theoretical perspectives set out in Chapters 2 to 4.

The observational study of the judgemental processes used in awarding was conducted at one examining board, focussing upon GCE A-level examinations. The observational study had three main phases. In the first phase, awarding meetings were observed and a category system for systematically recording their proceedings was developed. In the second and third phases, further observations were made and the category system was used to gather systematic data relevant to the above aims. The Phase 1 observations were made in Summer 1990. The second phase was carried out in Summer 1991 and the third phase in Summer 1993.

It was only possible to gather observational data during the summer of any given year because most awarding only takes place at that time. This imposed a somewhat lengthy timescale upon the study. (Although the board operates some examinations in the winter of each year, many fewer candidates and subjects are involved and the relatively small scale of the enterprise enables the awarding process for all subjects to be compressed into a few days, making it impossible to observe more than one or two meetings.) In addition, for practical reasons unconnected with the study itself, it was impossible to make any observations of awarding meetings in Summer 1992.

5.2 METHODOLOGY AND SCOPE OF THIS CHAPTER

5.2.1 Phase 1 Methodology

Phase 1 of the present observational work used *participant observation* (see, for example, Bogdan and Biklen; 1982) to develop the categorical coding scheme for use in later *systematic observation* (Bakeman and Gottman; 1986) in Phases 2 and 3. At the start of Phase 1, the awarding process was understood by reference to three sources of information:

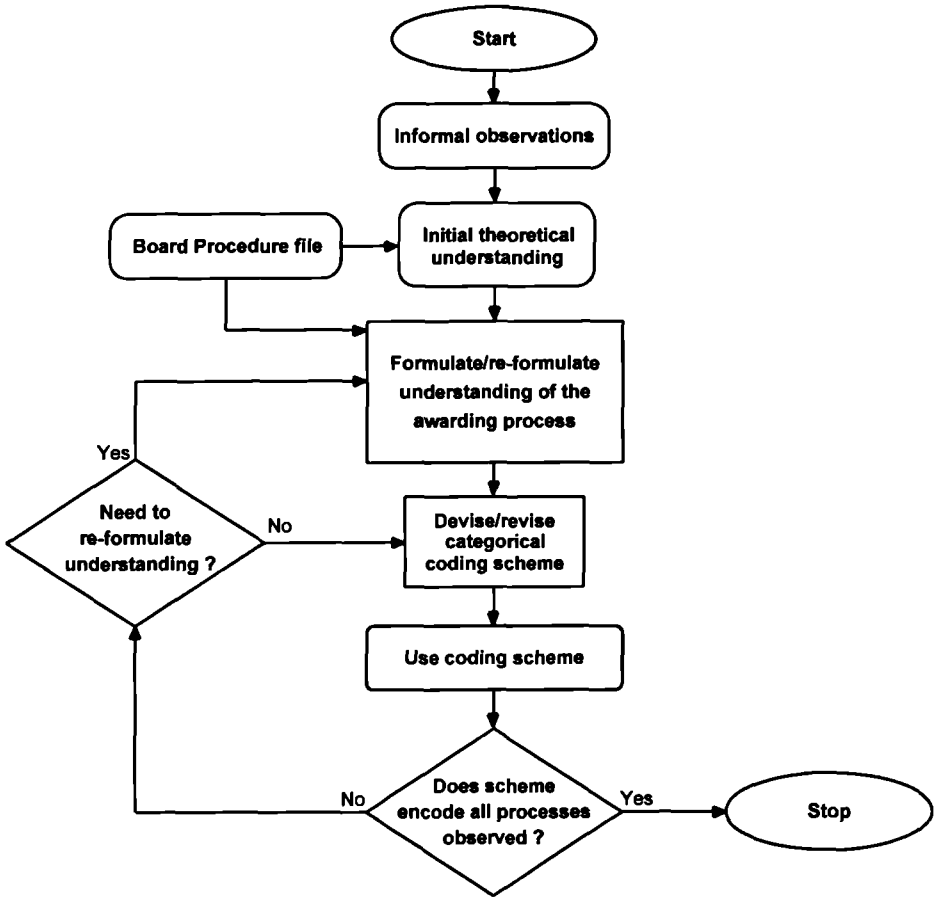
1 the board's 1990/1991 procedure document (see Appendix 5.1) which set out, for those involved in the process, the awarding procedures to be followed;

2 the author's previous informal observations as a participant in awarding meetings;

3 the initial theoretical identification of the awarding process as evaluative which is set out in Chapter 3 and grounded in those informal observations.

On the basis of the understanding of the awarding process provided by these sources of information, an initial coding scheme for systematic observation was constructed. Further observation and trial use of the coding scheme led, by a process involving elements of both *grounded theorizing* and *analytic induction* (Znaniecki, 1934; Glaser and Strauss, 1967; Bogdan and Biklen, 1982; Hammersley, 1989) to the simultaneous development of the understanding of the awarding process and the coding scheme. Figure 5.1 is an adaptation of Hammersley's Figure 1 (Hammersley, 1989; Page 170) illustrating the methodology used in Phase 1.

Figure 5.1 Methodology used to develop the observational categories
(after Hammersley, 1989)



From Figure 5.1, it can be seen that the methodology used in Phase 1 involved both the initial theoretical formulation which is characteristic of analytic induction and the categorical approach of the constant comparative method of grounded theory development. The sampling of awarding meetings during Phase 1 was purposive with additional meetings being observed until the developing categorical coding scheme appeared successfully to encode all the significant processes observed.

The identification of significant processes was, of course, a key factor in the development of the coding scheme. During Phase 1, modifications of the categories were grounded in the developing understanding of the awarding process, in the way described by Bogdan and Biklen (1982) and many other qualitative methodologists. Thus, it is not claimed that the categories in the final coding scheme reflect every detail of an objectively occurring set of

events. Rather, they encode those features of the awarding process which are significant within the understanding developed by the author during Phase 1 and from the theoretical perspectives discussed in Chapters 2 to 4. Certainly, other investigators with different understanding or theoretical perspectives upon awarding might choose to emphasise different features of the process in any coding scheme of their own. This is one of the main reasons why the contents of Chapters 2 to 4, above, have been presented at length before this one. It is also why later sections of this chapter report the simultaneous development of both the coding scheme and the author's understanding of the awarding process. It is hoped, as Becker (1970) argues, that this will enable the reader to follow some of the detail of these developments and thus reach a more informed judgement about the confidence which can be attached to the qualitative conclusions drawn.

5.2.1.1 The meetings observed in Phase 1

The Phase 1 work involved observing nine awarding meetings, two of which extended over two days. The subjects involved, in the order in which the meetings occurred, were:

1. Computer Science,
2. Sociology,
3. Physics,
4. English,
5. French,
6. Mathematics,
7. Chemistry,
8. Communication Studies,
9. Design and Technology.

The meetings were audio tape recorded to facilitate later qualitative analysis and a micro-computer program was written and used to encode the proceedings of the meetings in terms of the developing coding scheme. (The final version of this program can be found in Appendix 6.1.)

5.2.2 Methodology for Phases 2 and 3

Phases 2 and 3 of the observational study primarily involved the systematic application of the coding scheme developed during Phase 1, with the aim of building an accurate description of the practice of awarders. This aspect of the methodology of Phases 2 and 3 is described in Chapter 6 where the quantitative data from the systematic observations are reported.

However, qualitative data were also gathered during the Phase 2 and Phase 3 observations and are reported in this chapter.

5.2.2.1 The meetings observed in Phases 2 and 3

In the Phase 2 work, seven awarding meetings were observed, all of which lasted for a single day. The subjects involved, in the order in which the meetings occurred, were:

1. General Studies,
2. Accounting,
3. Physics,
4. Economic and Social History,
5. Mathematics,
6. English Language and Literature,
7. Communication Studies.

In Phase 3, five one-day meetings were observed. The subjects involved, in the order in which the meetings occurred, were:

1. Economics,
2. Physics,
3. Mathematics,
4. English Language and Literature,
5. Communication Studies.

In Phases 2 and 3, the observed meetings were chosen so as to represent the range of academic subjects examined at A-level. As for Phase 1, the meetings were audio tape recorded to facilitate later analysis.

5.2.3 The scope of this chapter

It is essential to an understanding of this chapter to be aware that significant changes were made to the board's awarding procedures between the Phase 2 and Phase 3 observations. These changes were largely the result of the Phase 1 and 2 work (both qualitative, as reported in this chapter, and quantitative, as reported in Chapter 6) which revealed a number of problems with the existing practice. This chapter therefore reports the detailed qualitative analysis of the awarding process, based upon observations made in the first and second phases of the work, which led to the changes in procedure. The changes are then described and their qualitative effects, as observed in Phase 3, are reported. (Practical difficulties related to the timing of meetings and the need for the author to attend certain other meetings

as a participant meant that it was not possible for awards in exactly the same subjects to be observed in Phase 3 as had been observed in Phase 2).

Subsequently, Chapter 6 reports the quantitative analysis of the systematically coded observations made during Phases 2 and 3 and further explores the effects of the procedural changes made.

5.3 THE BOARD'S AWARDING PROCEDURES DURING PHASES 1 AND 2 OF THE STUDY

This section sets out the procedures used at the board during Phases 1 and 2 (Summer 1990 and Summer 1991), with commentary based upon the informal observations which preceded Phase 1. The modifications to awarding procedures which were introduced prior to the Phase 3 observations in Summer 1993 are described in Section 5.6.

In each subject at A-level, the awarding meetings studied in 1990 and 1991 were required to make decisions about the minimum number of marks which should qualify candidates for each of three grades: A, B and E. At each grade, these decisions were made separately for each examination component and were then combined to give a decision for the examination as a whole. The overall examination boundaries at Grades A, B and E were then ratified by considering all the components together. (There was a national agreement between all the GCE boards specifying how the remaining grade boundaries should be interpolated along the total mark scale to give grades of approximately equal width; see Appendix 5.1 for details. This agreement is now part of the code of practice governing A-level examinations; SCAA,1994)

The participants in the awarding meetings consisted of the Chief and other senior examiners for the examination concerned, members of the board's Standing Advisory Committee (SAC) for the subject concerned and the board's Subject Officer for the examination. Reference to Appendix 5.1 shows that the formal responsibility for the decisions of the meeting lay with the SAC members, alone. The role of the examiners and board staff was to report and advise.

5.3.1 Preliminary reports - Step 0

In the board's 1990/1 procedure, each awarding meeting began the consideration of each component with a report from the Chief Examiner on the "response of candidates to it" (Appendix 5.1). This preliminary stage of the meetings will be called Step 0 (zero). At the end of his or her report, the Chief Examiner recommended particular marks as possible boundaries for Grades A, B and E on the component concerned. Some meetings began with a report on all components from the Chief Examiner, rather than beginning the consideration of each component with such a report. This tended to occur when the Chief Examiner was responsible for setting and coordinating the marking of all the components. In examinations where a team of senior examiners each had responsibility for a particular component, their individual reports and recommendations were generally taken immediately prior to the consideration of their component. No statistics from the examination were discussed at this stage. Following the examiner's preliminary report and recommendations, the meeting turned its attention to determining component grade boundaries. For each grade boundary decision, two procedural steps could be identified for each component:

1 scrutiny of individual scripts;

2 decision about the grade boundary for the component.

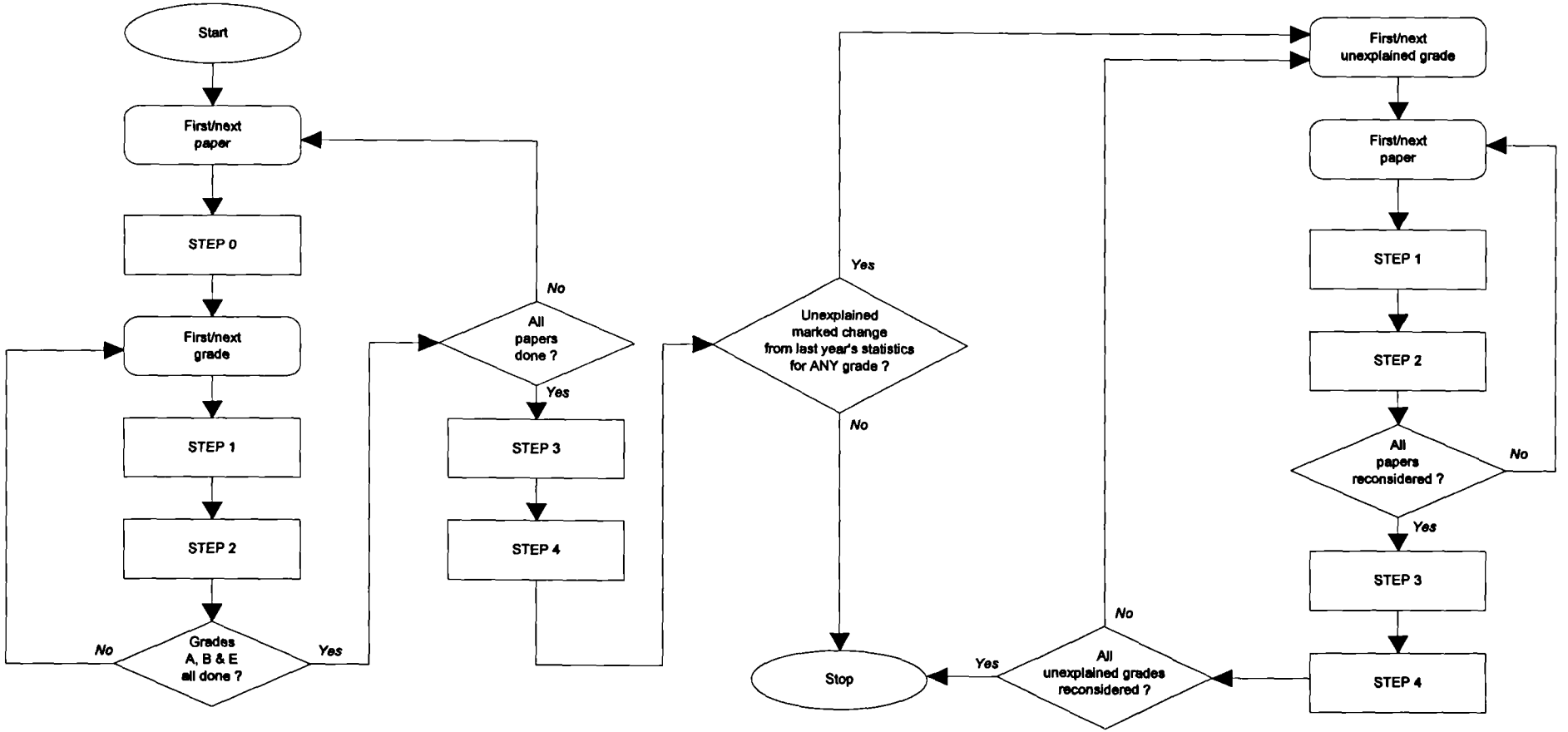
Once all the decisions had been made at a given boundary for each component, there were two further steps:

3 combination of the component decisions to produce a boundary for the examination as a whole;

4 ratification of the examination boundary.

There was considerable variation between meetings in the extent to which the examiners participated in these four steps. In meetings at one extreme, they took a full part in each step; at the other extreme, the examiners made no contribution beyond their initial report and recommendations, sometimes even leaving the meeting once their report had been received by the SAC members. Figure 5.2 shows the awarding procedures which were in use during Phases 1 and 2.

Figure 5.2 The awarding procedures in use during Phases 1 and 2 (1990/1)



5.3.2 Step 1 - scrutiny of individual scripts

In Step 1, the primary activity was the classification of individual scripts into one or other of the two grades on either side of the boundary being considered. The declared purpose of this activity was to enable each awarder to form a view of the number of marks awarded to scripts which he or she considered to be just worthy (and not quite worthy) of the upper grade. The awarders normally worked as individuals but sometimes referred detailed points to one or two physically nearby colleagues for their opinion. The instructions given to awarders (see Appendix 5.1) were not precise enough to lead to a common practice during Step 1. Some awarders attempted to make holistic judgements of the quality of each script but others appeared to concentrate on particular aspects (sometimes particular questions) and to make their judgement principally on this evidence. It was not uncommon to observe an individual awarder oscillating between these two ways of working, although this might have been simply part of the process of forming an holistic judgement.

5.3.3 Step 2 - decision about the grade boundary for the component

Step 2 was done collectively and relatively formally with the chairperson organising the discussion. In general, each awarder was given an opportunity to express a view on where the boundary should be positioned and any initial disagreement led to discussion of scripts scrutinised in Step 1 until a consensus was reached about the boundary mark. In the meetings observed, the discussion always produced a consensus but there was provision for voting in the event of a deadlock. During discussion in Step 2, all the considerations which were brought to bear to judge individual scripts during Step 1 could be discussed.

During Step 2, statistical comparisons were sometimes made with the grade distributions produced by the decisions for the same boundary upon different components and, more often, comparisons with the grade distributions from the component in question in previous examinations in the same series. These comparisons were sometimes initiated by the Subject Officer and sometimes by the awarders. During these statistical comparisons, reference was sometimes made to the examiners' reported (Step 0) impressions of the general quality of the work produced by the current candidates as a group, particularly in comparison with that produced in previous examinations.

5.3.4 Step 3 - combination of the component decisions to produce a boundary for the examination as a whole

Step 3 was a mechanical process in which, for each grade, the boundary marks established for each component were added together by the Subject Officer (using any scaling factors which were applied to the examination during the totalling of candidates' marks) to produce boundaries upon the aggregated mark scale of the examination as a whole. Thus, the board's 1990/1 procedures used the Model 1 addition method set out in Chapter 4. In this connection, the board's procedure file (Appendix 5.1) notes that:

> "...regression-to-the-mean effects, arising from imperfect inter-paper correlations will result in cumulative percentages of candidates at the overall grade boundaries which are, in general, different from those at the component boundaries."

5.3.5 Step 4 - ratification of the examination boundary

Step 4 had two aspects. One was the consideration of the grade distributions produced by the examination boundaries, both in comparison with previous examinations in the series and in the light of examiners' impressions of the general quality of the current candidates' work. The other aspect was further scrutiny of all the work of candidates at the boundaries. However, this latter process was normally carried out only when major changes from the previous year in the grade distribution were implied by the original judgements. In most cases, ratification only involved reviewing the statistical information.

Step 4 was a mixture of formal and informal proceedings. The consideration of the statistics was done by the group as a whole, advised (to an extent which varied considerably between meetings) by the Subject Officer. In some meetings, one of the board's Research Officers or a member of the senior administrative staff was asked, by the Subject Officer, to attend at this stage. This occurred when the Subject Officer felt that the grade distribution produced by the unratified boundaries represented a "marked" (Appendix 5.1) change, which the awarders could not explain, from the previous year. Further scrutiny of candidates' work, when it took place, was usually done individually or in small *ad hoc* sub-groups of two or three. Thus, ratification of the overall subject boundaries in Step 4 could involve both statistical data and the same sort of qualitative judgemental data as Steps 1 and 2. The

weight given to these two different types of information appeared to vary considerably between meetings.

5.4 THE INITIAL OBSERVATIONAL CATEGORIES

Since the participants in awarding meetings work collectively to determine grade boundaries, the social interactions between them might be expected to influence their decisions (Brown, 1988) and therefore appeared worthy of systematic study. With the exception of Christie and Forrest (1981), previous research on examination standards had not considered this aspect of the awarding process at all. Moreover, Christie and Forrest considered only the relationship (which they identified as a "contest") between examiners and the examining board secretariat. One group of observational categories which it was thought would be appropriate to try, therefore, were ones which encoded the interactions between the participants in the awarding meeting. These will be referred to as the *dynamic* categories and are concerned with the socio-psychological interactive processes within the awarding meeting. Since the fieldwork described in this chapter was carried out, a study of the dynamics of grade awarding for GCSE examinations has been done by researchers at *Nottingham University* for the *Schools Curriculum and Assessment Authority* (Murphy *et al*, 1996). The Nottingham study confirmed the importance of the social aspects of awarding meetings.

However, the principal focus of the present study is on the substantive content of awarding meetings and, as outlined in Section 5.3, two kinds of evidence are used by awarders to reach their decisions: candidates' scripts and statistical information (from the current examination and from previous examinations in the same series). It was therefore necessary to attempt to encode both of these facets of the judgemental process of the meetings; the relevant category groupings will be referred to as *evaluative* and *statistical* respectively.

5.4.1 Dynamic categories

In the terms of social psychology, awarding meetings can be identified as small decision-making groups and considerable work has been done on how such groups operate (Brown, 1988; Baron, *et al* 1992). One of the most widely used systems of categories for encoding social interaction within small groups was developed by Bales (1950, 1970). His system involves twelve categories, for each of which Bales has developed an extended description so as to standardize usage. The categories are:

1 Seems Friendly
2 Dramatizes
3 Agrees
4 Gives Suggestion
5 Gives Opinion
6 Gives Information
7 Asks for Information
8 Asks for Opinion
9 Asks for Suggestion
10 Disagrees
11 Shows Tension
12 Seems Unfriendly.

Note that most published descriptions of Bales' categories refer to the original 1950 formulations whereas the above are the categories as revised by Bales on the basis of extensive empirical work (Bales, 1970). Bales' categories can be cross-classified in two ways. Categories 1 to 3 and Categories 10 to 12 form two subgroups which refer to the affective content of interactions; Categories 4 to 6 and Categories 7 to 9 form two subgroups which refer to types of interactions focussed upon the task which the group is addressing. Considered in pairs, (1 and 12, 2 and 11, and so on) the categories reflect opposite aspects of the same facet of interaction.

5.4.2 Evaluative categories

To provide a systematic categorisation of the evaluative aspects of awarding, studies of similar processes in other fields were considered. As noted in Chapter 3, a field which has obviously similar concerns is aesthetics and the reasons which might be given to support evaluative aesthetic judgements offer a useful taxonomy for categorising the types of judgements made by awarders and for considering the legitimacy of those judgements. Beardsley (1981) distinguishes five different types of reasons which might be given to support aesthetic judgements. Modifying his approach and terminology slightly, it is possible

to identify several different types of reason which could be given to support an evaluation of examination candidates' work.

5.4.2.1 Genetic reasons

Reasons in the first group, known as *genetic* reasons, refer to the manner in which the work was produced, the intention behind its production or its relationship to other work, for example:

It achieves its aims fully.
It is skilfully presented.
It is illegible.
It is original.
It is derivative.

Consider the first of these reasons, which refers to the pupil's intention. In aesthetics, there is a long running debate concerning the appropriateness of judging a work against the artist's intention (for the two points of view, see Savile, 1972 and Beardsley, 1981). The debate centres around two main questions: can the finer nuances of the artist's intention (which may, in any case, be modified during the production of the work) ever be known with certainty? Secondly, is the aim of criticism to evaluate the work itself? If so, the artist's intention is irrelevant because even if it is completely achieved, there is still a need to evaluate whether it was worth achieving. Taking up this second point, Beardsley (1981) argues strongly that, as a result, considerations of intention, and the other genetic reasons including originality, refer to the artist but not the work itself, although he qualifies his inclusion of skill in this category to accommodate Stolnitz (1973) who argues that skilfulness may be appreciated as a characteristic of the work per se (that is, as an *objective* reason; see below).

When the genetic reasons are considered as reasons to support an evaluation of a pupil's examination work, however, the questions which are debated in aesthetics become much less controversial. Firstly, the work which is being evaluated during the grading process is produced as a response to a specific task and the intention of the pupil is therefore usually well understood. Secondly, the purpose of the examination is to evaluate the pupil (or, rather, his or her attainment) so that if, as Beardsley (1981) argues, genetic reasons relate to the pupil, rather than the work *per se*, their use appears entirely appropriate. Indeed, it is

those genetic reasons which most obviously describe the work itself which are most controversial as reasons given in support of an educational evaluation of the type involved in awarding. The example of legibility, given above, is a case in point. Some would argue for legibility to be a criterion in all assessments, others that it is a legitimate basis for evaluation of educational attainment only in a few specialised contexts.

5.4.2.2 Moral and Social reasons

Reasons in the second group refer to the *moral and social* aspects of pupils' work, for example:

It is amoral.
It is effective social criticism.
It is subversive.
It is morally uplifting.

The legitimacy of these reasons in the context of setting examination grading standards seems doubtful. It is possible to envisage contexts in which the second example refers to a quality which is relevant to an evaluation of pupils' work but this will usually reflect a concern for the cognitive aspects of the social criticism offered. On the other hand, suppose that pupils' ability to empathise with people in historically remote times was an aspect of attainment which was to be evaluated. Suppose, further, that the pupils had been asked to empathise with Sir Thomas More's opposition to Henry VIII's divorce from Catharine of Aragon. Would not the moral values in the work they produce then be relevant to its evaluation? The other obvious case in which moral features of pupils' work could be relevant is in examinations of Religious Education. However, these are special cases and, in general, the moral and social aspects of pupils' work are not relevant to the awarding process.

5.4.2.3 Affective reasons

The third group of reasons which might support grading judgements refer to the psychological effects of the pupils' work on the evaluator. *Affective reasons* include the following:

It gives pleasure (I enjoyed reading it).
It is interesting (it interested me).
It is impressive (it impressed me).
It is tedious (it did not excite me at all).
It is pathetic (it aroused my sympathy for the pupil's lack of attainment).

In general, affective reasons do not seem to be legitimate reasons to give in support of an educational evaluation (although they may be relevant in the specialised circumstances of the assessment of artistic and creative work where the ability to produce such effects in an audience may be a valued skill). This is because affective reasons generally beg the question *why*? If an awarder finds a pupil's work interesting or impressive, then he or she can reasonably be asked to identify those features of the work which give it this quality. Although the awarder will not, necessarily, be able to give an answer, since providing convincing explanations for affective responses is notoriously difficult, if an answer is given, it will fall into one of the other categories outlined in this section.

5.4.2.4 Objective reasons (Unity, Structure, Complexity, Intensity)

Beardsley calls the fourth type of reasons for evaluative judgements which we shall consider, *objective* reasons and divides it into three sub-types concerned with the *unity and structure* of the work, the *complexity* of the work and the *intensity* of the human qualities in the work. The following are examples of reasons referring to unity and structure:

It is well organised.
It is chaotic, lacking structure.
It has stylistic unity.

The following are examples of reasons referring to complexity:

It is rich in contrasts.
It is subtle and detailed.
It is simplistic.

Finally, reasons relating to intensity are exemplified by the following:

It is forcefully argued.
It is lively.
It has delicacy.

Reasons relating to unity and complexity seem to have a general legitimacy in the context of educational evaluations. Those which refer to intensity are more problematic since the distinction between this sub-type and affective reasons is not always clear and it seems possible to attack reasons of intensity in a similar way. For example, it seems reasonable to ask an evaluator what features of a piece of work make the argument it contains forceful. Is this a distinct quality which cannot be accounted for in terms of unity, complexity and content? It seems unlikely.

5.4.2.5 Objective reasons (Content)

Mention of *content* introduces the final group of reasons which, according to Beardsley, might be offered in support of a particular evaluative judgement. He identifies this group as referring to the cognitive aspects of a work as in the following examples:

It has something important to say.
It is a competent review of the issues.
It is an effective solution to the problem.
It does not advance the argument.
It omits several important points.
It is a partial solution to the problem.

These reasons are concerned with the quality of the intellectual content of work. As such they are legitimate reasons to use to support a particular evaluation of educational attainment even if, as Beardsley argues, they are not appropriate in a purely aesthetic context. They can be seen as a further sub-type of objective reasons.

5.4.2.6 The initial evaluative categories

The evaluative categories which were tried initially were developed from the classification summarised above. The categories were as follows:

1 Genetic, Moral, Social and Affective
2 Objective - Unity, Structure and Complexity
3 Objective - Content.

5.4.2.7 To what do the reasons refer?

From the description of Step 1 of the awarding process given in Section 5.3, it will be clear that the reasons given by awarders in support of their evaluations of individual scripts might reflect an holistic consideration of a candidate's work or might apply only to some particular aspects of the script. These different approaches to the evaluative task were called *holistic* and *fragmented* respectively and it was decided to record them separately.

In Chapter 3, the importance to the awarding process of the context in which candidates perform was emphasised and it was concluded that the maintenance of standards requires awarders to judge candidates' work as responses to particular tasks. However, from informal observations, awarders appeared rarely to concern themselves **explicitly** with context in the widest sense. The major artificialities of the performances required in public examinations (for example, the formality of examination conditions, the time pressures on

candidates, the type of examination component concerned and so on) were either ignored or taken for granted as part of the examination scenery.

However, some awarders did appear to address narrower contextual issues such as the difficulty of particular tasks. It was therefore decided to attempt to record evaluations made in the observed awarding meetings as coming either from a *contextualised* or *uncontextualised* judgemental process. (Note that the prefix *un* has been chosen deliberately. It is judgements which are made **without reference** to context which fall into this category, not judgements which make allowance for context effects and might therefore be called *de*-contextualised).

Examples of the different types of evaluation which it was hoped to identify are:

> *This isn't a Grade B script, he can't integrate a function of a function.*
> (Fragmented, Uncontextualised).
>
> *This is a Grade B script, she did well on that tricky integral in Question 10.*
> (Fragmented, Contextualised).
>
> *This is a Grade B script, she didn't get the integral in Question 10 out but its a tricky one and the rest of the work is good.*
> (Holistic, Contextualised)
>
> *This isn't a Grade B script, in an open book examination like this you're looking for a much deeper level of analysis for Grade B.*
> (Holistic, Contextualised)
>
> *I don't think the work in this script is good enough for a Grade B. It's nowhere near as good as the Grade B coursework we were just looking at.*
> (Holistic, Uncontextualised)

It was anticipated that the application of these categories would be problematic because of the nature of the comments being classified. The statements made by awarders during their task might be incomplete since their purpose is not necessarily to communicate the exact thought processes being used. Much of their meaning might be implied. For example, the first instance given above might come from an awarder who has actually made allowance for the nature of the particular integral used in the examination. He or she might feel, with some justification, that the context of the remark would imply this to his intended audience who would be other awarders engaged on the same task. If this were so, then a contextualised

judgement would actually have been made. Similarly, an awarder making the second comment might not be referring to the candidate's performance on the integral in Question 10 as **the** justification for giving it a Grade B. Instead, the second part of the comment might simply be intended as an example of the many reasons why a Grade B is justified. If this were so, then the judgement in question should be classified as holistic, rather than fragmented. Related difficulties concern the presence or absence of qualifying adjectives. For example, if an awarder does not explicitly refer to a task as difficult when assessing a candidate's response to it, does he or she thereby imply that it is easy or that its difficulty is irrelevant to the awarding task? Again, in the final example above, the awarder does not indicate in the first sentence **why** the script is not Grade B so it is assumed that this is an holistic judgement. However, the judgement might be the result of an **unstated** observation that "he can't integrate a function of a function" in which case it should be classed as fragmented (see the first example).

A second class of problems arise because it may not be the case that the awarder is actually making judgements in the way he or she says. It may, for example, be that features of the script which a particular awarder values are not ones which are widely held to be relevant and the awarder might not, therefore, refer to them in discussion, preferring to address less controversial issues in the hope of being more persuasive. Alternatively, it might be the case that awarders' judgements about grades are not formed as the result solely of an analytical process but are expressions of affective, as well as (or even instead of) cognitive response to the scripts. If this is so, then the sort of comments exemplified above must be seen as merely post hoc rationalisations or, at best, partial attempts to persuade others by reference to qualities in scripts which the awarder believes to be relevant to their cognitive evaluation.

Despite these potential problems, it was decided to attempt initially to classify the usage of the evaluative categories as follows:

1. Holistic or Fragmented
2. Contextualised or Uncontextualised

5.4.3 Statistical categories

The initial statistical categories were the least theoretically developed of the three groupings. However, on the basis of informal observation of awarding meetings prior to Phase 1, it was decided to adopt the following two categories:

1. Mark orientated;
2. Grade orientated.

The distinction between mark and grade orientation reflects the fact that sometimes comparisons are made between the number of marks required for the grade in a previous year or on another component and sometimes the statistical comparison is with the grade distribution from a previous year or another component. Statistical reference in awarding meetings is usually comparative in nature and two approaches can be distinguished. The *historical* approach refers to comparison with previous examinations in the same series and the *internal* approach refers to comparison with other components of the current examination. It was decided to attempt the separate coding of these two approaches for each of the two statistical categories.

5.4.4 Merging the category groupings

The full coding scheme which formed the initial starting point for development work in Phase 1 of the study involved all three groupings, integrated into a single structure as follows:

1. Shows Positive Affect
2. Agrees
3. Gives Suggestion
4. Gives Evaluative Opinion
   - a. Genetic, Moral, Social and Affective
   - b. Objective - Unity, Structure and Complexity
   - c. Objective - Content.

   (Usage of these subcategories is either: (i) Contextualised or Uncontextualised; (ii) Holistic or Fragmented.)
5. Gives Statistical Opinion
   - a. Mark orientated
   - b. Grade orientated

   (Usage of these subcategories is either Historical or Internal.)
6. Gives Information
   - a. Statistical
   - b. Other
7. Asks for Information
   - a. Statistical
   - b. Other
8. Asks for Statistical Opinion
9. Asks for Evaluative Opinion
10. Asks for Suggestion
11. Disagrees
12. Shows Negative Affect

In the context of the dynamic categories, both the evaluative and the statistical categories are subcategories of Gives Opinion. Although it is theoretically true that Asks for Opinion could be sub-divided in the same way, it was not. There were two reasons for this: firstly, the number of categories to be coded was already large and problems of manageability were anticipated; secondly, from informal observation prior to Phase 1, it seemed that awarders' requests for evaluative opinion were not phrased so as to refer to one evaluative category but, rather, used formulations such as *what do you think of this one?*. On the other hand, opinion on statistical matters was sought distinctly from evaluative opinion so it seemed necessary to make at least this cruder distinction within Asks for Opinion.

It was apparent before the first formal observations in Phase 1 that the complexity and number of the categories proposed was likely to make coding the meetings extremely difficult. It was therefore an explicit aim of Phase 1 to find ways of simplifying and reducing the initial coding scheme. It was also decided to tape record some of the observed meetings and to try out the use of a micro-computer as an aid to data-capture.

5.5 QUALITATIVE ANALYSIS AND CATEGORY SYSTEM DEVELOPMENT

As explained in Section 5.2, the Phase 1 work was characterised by continuous development of understanding of the awarding process and modification of the categories in the coding scheme in the light of experience. In this section, the features of the Phase 1 meetings which had a major role in developing understanding of the awarding process and the contents and use of the coding scheme for Phases 2 and 3 of the study are described and analysed. The qualitative analysis reported in this section also draws upon evidence from the meetings observed during Phase 2.

5.5.1 Step 0 - preliminary reports

It became apparent during the observational work that the nature of the preliminary reports given to the awarding meetings was of considerable significance. The Chief Examiners' reports generally followed the brief given in the board's procedure paper (Appendix 5.1) by describing how the candidates had responded to the current year's paper(s). Thus, remarks

such as *Candidates did not do as well this year as last* and *In terms of calibre, I would say slightly, not much, better quality of candidates this year* were typical. These reports tended, therefore, to focus subsequent discussion upon the quality of the candidates' attainments and not upon the way in which the current year's papers had operated. Thus, they were more frequently followed with speculation about why the current candidates were better or worse than previous ones, rather than whether the current question paper was harder or easier than previous ones. Indeed, there was a general tendency, exemplified by the following extract from the Accounting meeting, to defend questions which had proved unexpectedly easy or difficult by attributing this to the candidates, rather than to the questions.

CE[*] *Question 1, the nice easy opener, turns out to be a killer of course... as is our tradition on these papers... Answers to Part A were rather better than Part B. Many candidates, particularly average sort of candidates really didn't understand what was meant by B and found some difficulty in taking a figure from their control accounting and attempting to work back to anything. It was the better candidate who was able really to approach Part B at all. So Question 1 - not particularly well done.*

Chair *Did you find that any were getting into the top register, getting maximum marks?*

CE *Yes, I did see some maximum marks but it was literally, you know, a handful of candidates who achieved that.*

Chair *But the weaker candidates only did Part A?*

CE *Yes, quite a few, particularly the weaker ones.*

A1[♦] *They had heard of a control account somehow but...?*

Chair *I think the problem with that sort of question, and the reason it is such a good question, is that it actually tests knowledge of double entry so even if you know what a control account is, it doesn't actually help you with the answer. It's a starting point but you have to understand the double entry system and that's what candidates found difficult. They know what a control account is. If you ask them just to do a control account they could probably do one but set in the context of this type of question they find it very demanding. It* ***is*** *a very good question.*

CE *It has some important practical connotations.*

Chair *Absolutely... Right, so, any questions? Come in and ask questions whenever you want.*

A2 *I'm sorry... with Question 1... it is a bit of a* ***down*** *in terms of responses. Is there any problem with the wording or anything like that? I was just thinking of the future, you know.*

CE *Well I think any question which taxes candidates' overall knowledge of the double entry bookkeeping system they will find difficult. My own view is that it is a perfectly valid question to ask in that they are questions we have sometimes asked over the years and I would have thought that as a matter of policy it was the sort of question we would want to continue to*

[*] CE = Chief Examiner

[♦] A1 = First awarder; A2 = Second awarder; etc.

include because it is one of the vehicles of testing knowledge of double entry.

Chair *I think that there aren't enough questions of that type*

CE *Yes*

Chair *There's nothing tricky about it.*

CE *No, no.*

Chair *You see, its your own background. You actually learned doing control accounts properly, for real. You see, most candidates don't.*

A2 *No, I suppose they don't and they have to rely on doing questions, which in a sense are artificial...*

Chair *What is a bit worrying I think is that we thought that this was an easy one - I must say I did - it is basic and it is foundation knowledge and if A-level candidates can't get this right its very difficult to see how they are worth much.*

CE *That's true.*

Chair *Right, shall we move on to Question 2?*

This pattern of argument was observed very frequently during awarding Step 0 and is not only irrelevant but also positively misleading from the point of view of the maintenance of awarding standards from one year to the next. In Chapter 3, it was argued that, to maintain standards, awarding judgements must be contextualised by use of the *script as response* strategy. To be consistent with this strategy, the issue which the Accounting awarders should have addressed was whether the question in this year's paper was more difficult for candidates than the corresponding question in last year's, not whether it was a legitimate question. However, when Awarder 2 began, somewhat apologetically, to explore the issue of the specific difficulty of Question 1, she was answered with an explanation of why the candidates' attainment was poor. This explanation may well be correct but is not relevant to the task of maintaining standards unless it was being argued that the weakness identified in this year's candidates was not present among last year's. This was clearly not the argument being put; indeed it was strongly implied that candidates would have scored badly if the present Question 1 had been set in any recent year. Thus, Question 1 must have depressed the scores of this year's candidates relative to last year's and the *script as response* strategy (Section 3.5.1) implies that the grade boundaries, on this evidence at least, should therefore be set slightly lower than in the previous year when no such question was set. The accounting awarders' discussion (especially the penultimate quoted remark from the Chair) implies, instead, the use of the *script as artefact* strategy and no consequent reduction in the number of marks required for each grade.

It is interesting to reflect upon why awarders in general, for the above type of discussion was common during Step 0, were predisposed to ascribe unexpected apparent difficulty or easiness to the candidates, rather than the questions. It seems probable that, to some extent at least, this occurred because the awarders were aware of their own responsibility for the questions. The Chief Examiner had a clear responsibility because he or she set the questions originally. However, many of the awarders, especially the Chair, had also been involved in scrutinising and amending the draft question papers, sometimes considerably. The awarders were likely, therefore, to feel that they had already considered the difficulty of the questions in detail and satisfied themselves that it was appropriate. A reluctance to revise this view seems natural although, as noted in Chapter 3, predicting the difficulty of examination questions is, in fact, extremely difficult (see, also, Cresswell, 1994).

It must be noted that, formally, it is impossible to determine, on the basis only of the current year's candidates' responses to the current year's questions, to what extent the candidates' attainments are differently distributed from those of the previous year and, on the other hand, to what extent the questions are easier or more difficult. However, it is necessary to do this if comparable standards are to be maintained between the two years (see Chapter 3) and awarding procedures assume that awarders' implicit judgements about the difficulty of the questions are the additional data required to enable standards to be maintained. The scrutiny of candidates' scripts, it is argued, enables the awarders to improve their original judgements about the difficulty of the questions *vis-a-vis* those of the previous year. Unfortunately, there was little observable evidence that awarders interpreted candidates' scripts in this way during the preliminary discussion in Step 0 or, as will become clear, at any later stage of the awarding meeting.

In this connection, the board's policy about the provision of statistical data to the awarders is relevant (see Appendix 5.1). During Phases 1 and 2 of the study (Summer 1990 and 1991) statistical data from the examination were deliberately withheld from the awarders until Step 2 of the awarding process and were not discussed in any detail or interpreted until Step 4. The rationale for this policy was the desire to prevent statistical considerations from affecting the awarders' qualitative evaluation of the scripts. However, as reported in the following

sections, this policy did not prevent awarders, in Steps 1 and 2, from bringing to bear considerations other than their professional judgement of the quality of candidates' work. It seemed reasonable to suggest that the board's policy about the early provision of statistical data to the awarders should be reviewed. For example, a statistical report during Step 0, alongside the qualitative one provided by the Chief Examiner, might have served as a useful vehicle with which to focus the Accounting meeting's attention upon the question of how the demands made by the current year's papers compared with those of the previous year's. This issue will be discussed again in the context of Step 4 of the awarding process (Section 5.5.5) and in Section 5.6.

Although there were variations between the observed meetings in the amount of discussion of the papers and candidates in general which occurred, principally during Step 0 but also at other times, substantial amounts of time were spent on these matters by all the meetings. As this was discussion of the context within which the candidates had performed, it was clearly necessary to accomodate it within the coding scheme so two new categories were introduced for this purpose during Phase 1.

5.5.2 Step 1 - scrutiny of individual scripts

During Phase 1 of the study, it rapidly became apparent that, in awarding Step 1, the nature of the meetings varied considerably from subject to subject and made encoding the proceedings extremely difficult. As far as Step 1 of the meetings was concerned, two of the nine Phase 1 meetings scrutinised the scripts in almost complete silence, in one of the other meetings there was continuous noisy discussion between awarders and in the remainder something in between these two extremes occurred. Moreover, awarders had no reason to vocalise their thoughts at all except when consulting a colleague. In response to this, the reduced aim was initially adopted of simply noting the frequency with which the various evaluative categories were mentioned during the informal discussions in Step 1 without attempting to ascribe their use to individual participants. However, even this proved impossible because of the frequency, inaudibility, simultaneity and, more fundamentally, allusiveness of discussions between awarders in Step 1. For example, the comment *look at*

this, particularly that bit there was answered simply by a smile and a nod. It is extremely difficult to encode such an interaction. Was the first awarder giving or asking for an evaluative opinion? Was the second awarder giving an evaluative opinion, expressing general agreement or simply showing positive affect? Were the implicit judgements contextualised or uncontextualised? Inevitably, the comments of a pair of awarders informally discussing a particular script are difficult to interpret unless the script itself can be studied at the same time. As a result of these difficulties, it was decided to issue awarders in Phase 2 of the study with forms on which they could record their reasons for the evaluations made of individual scripts. These reasons would then be analysed using the evaluative categories. This approach proved more successful and the results are reported in Chapter 6.

Nonetheless, in terms of the original coding scheme (Section 5.4.2), both affective and objective reasons were mentioned during most of the meetings and both holistic and fragmented judgements were observed. The brief given to the awarders in the board's procedure file (Appendix 5.1) did not lead to a common approach to the judgement of scripts during Step 1. Even within a single meeting, some awarders appeared to make holistic judgements of the quality of each script whereas others concentrated on particular aspects (sometimes particular questions) and appeared to make fragmented judgements principally on this evidence. Comments such as this, from one of the Physics awarders: *You can't give a pass in Physics to someone who gets nothing on that question* occurred at some point in many of the observed meetings. Murphy *et al* (1996) report similar findings for GCSE awards.

During Step 1, several of the observed meetings spent a little time discussing the criteria which they should use to decide whether scripts were at the particular grade boundary. When it occurred, this discussion was not profound, rarely going beyond general and loose description of the characteristics of work worthy of the grade in question. Nonetheless, such discussions indicated that the awarders were attempting to evaluate the content of the scripts rather than simply responding affectively. A category to encode such discussion of evaluative criteria was introduced into the coding scheme.

It was also clear from the conversation between some awarders that they brought considerations other than their qualitative judgement of individual scripts to bear during Step 1. These were primarily historical statistical considerations, involving the proportion of marks which the script had earned and whether this was similar to the proportion of marks earned by borderline scripts on previous occasions. The arbitrary nature of mark scales (see Chapter 3) was generally acknowledged in all the observed meetings (for example, the following interchange occurred in the Communication Studies award: *I think this 24 is a pass but it was 26 last year. But it* [the Grade E boundary] *wont be the same every year.*). However, most awarders also clearly felt that the degree of arbitrariness was limited and that any particular grade boundary would normally lie reasonably close, in mark terms, to its position in the previous year's examination. There were also occasional comments which implied the existence of some sort of absolute mark criteria. For example, *80% ought to get you a Grade A.*

Such views, particularly since they appeared to be shared by Chief Examiners, provide a further reason for introducing statistical data during Step 0 of the awarding procedures. In particular, such data could be used in the selection of scripts for scrutiny in Step 1. At the time of Phases 1 and 2 of the study, the Chief Examiner's recommendations were used as the sole basis for selecting scripts to be scrutinised. Since the observed awarding meetings rarely considered scripts other than those initially provided, the selection of scripts at appropriate points in the mark scale is clearly very important. If a paper is unexpectedly easy or difficult compared with the previous year, then it is necessary, under the *Script as Response* strategy, to peruse scripts at higher or lower marks, respectively, than last year's boundaries. Chief Examiners' recommendations did not, in the main, take account of such factors. These issues are considered in more detail, and illustrated with an extract from one of the observed awarding meetings, in Section 5.5.5.

The dynamics of Step 1 were very often complex since individual awarders consulted with several different colleagues at different times. An initial attempt was made during the Phase 1 work to document these contacts because it was hypothesised that individuals who

worked closely together in Step 1 might behave in a similar way during Step 2. However, recording the network of contacts made within Step 1 proved impossible. Again, comments were too frequently made simultaneously and responses were sometimes directed apparently to the group as a whole, any of whom might reply. In each meeting observed, every awarder effectively discussed at least some points with every other. Only detailed analysis of video recordings of the meetings might enable the frequency and nature of contacts of every awarder to be identified.

5.5.3 Step 2 - decision about the grade boundary for the component

As with Step 1, Step 2 did not take the same form in all the meetings observed. In particular, some meetings (Mathematics was the clearest example) began Step 2 by formally asking each awarder for a proposed boundary for the grade in question. The subsequent discussion then focussed on reconciling the various proposals which had been made. When this strategy was adopted, there were informal attempts made not to ask the awarders for their proposals in the same order at every grade boundary. The reason for varying the order in which the awarders responded was evidently to avoid putting the same individual in the hot seat of responding first every time. However, there rapidly developed in meetings which adopted this approach a rueful acknowledgement that the awarders were consistently differing. Thus, when one of the later grade boundaries was considered in the Mathematics meeting, one awarder felt constrained to remark: *Well I was a few marks below that but I seem to be a bit generous today so I'll go along with 34.* Thus, the long-established (see, for example, Deutsch and Gerard, 1955) tendency for people to be more likely to conform to a group norm if deviation is easily identifiable appears to have operated in this instance. The recent study of GCSE awarding (Murphy *et al*, 1996) also observed group conformity effects of this type.

Even when, as in the majority of meetings observed, awarders were formally free to keep their views to themselves, the usual (see, for example, Baron *et al,* 1992) pressures to conform to, or resist, a working group's judgemental norm appeared to operate in Step 2 of awarding meetings. Dissenting views, when offered, were frequently accompanied by an

apologetic preamble and the status enjoyed by the Senior Examiners and Chairs of the meetings appeared to accompany a greater independence on their part. Both of these effects are evident in the extract from the Physics meeting given later in this section. Given that most of the observed meetings were conducted relatively informally and the awarders were peers who generally knew each other well and had done the same job together before, it seems reasonable to infer that the pressures to conform were mainly informational, rather than normative. However, the awarders in each subject (Senior Examiners and SAC members, collectively) clearly defined themselves as a social group whose membership they generally valued and, in these circumstances, there is also some normative pressure to conform, as Back (1951) observed.

The most surprising feature of Step 2 to emerge during Phase 1 of the observational study was the extreme rarity of specific evaluative comment. During the discussion in this step, it was common for awarders to refer to individual scripts and express views about the grade which they merited but reasons for these judgements in terms of particular features of the candidates' work were hardly ever offered. This was true even in those meetings which had briefly discussed evaluative criteria in Step 1. Thus, rather than being a debate in which reasons were given for particular evaluative judgements, discussion during Step 2 was a process of negotiating a compromise grade boundary which all the awarders felt able to support, or at least accept, given their personal positions based upon their own evaluations of the scripts. The following extract from the discussion during the setting of the Paper 2 E/N boundary in Physics is typical:

CE: [reporting in Step 0] *...What I do when I mark my scripts always is to produce an estimate grade as I mark it, my impression grade. And when I went through it at the end, the lowest one to whom I gave a questionnable E , and that's always a good sign you're on a borderline, was 49. So I'm going to go for 49.*

Chair: *49, thankyou Ken. Right, can I suggest that you all now look at the scripts between 45 and 52 which John* [the Subject Officer] *has got for us?*

[Step 1]

Chair: *Right. Are people settling in their minds where they'd like to be?* [pause] *It would be with the greatest reluctance and not altogether convinced that we weren't slightly easing things, if I were to accept a 49. Right, now, from all sides please.*

A1: *I think I could live with a 48 and even some lower... you can see there are things they can do.*

Chair: *Therefore, they're given near misses. That is not the issue. The issue is what mark this year, on this paper, corresponds as near as we can tell to what we did last year. If you were to look through the bottom end of the 40s, then you could say exactly the same: that here and there, there is stuff that they've got hold of.*

SO[*]*:* *Chairman, A2 marked Paper 2. Does he have a suggestion of where we should look?*

A2: *Well, I was thinking in terms of 50. That was a figure which I remember we mentioned on Monday* [at an examiners' meeting]. *But, certainly a lot of the scripts that I marked that were between 45 and 55... really you could pitch it anywhere between those and it would be...*

Chair: *Did you mark Paper 2 last year?*

A2: *Yes. Where did we put it last year?*

Chair: *47 last year. And as I said earlier, my impression is that if we pitch it at 49 this year we are being possibly a little kind.*

A3: *Well on looking through, having thought 50 by just looking at this year's paper, I've moved up to 51 to be honest.*

A4: *I've put down the 50s were just about in and the 49 I looked at had three quite good Section A answers and two reasonable Section B. The B and D Sections were a bit iffy. So I think the 50s are in and the 49, that I would say is doubtful. I'm not happy about the 48s.*

CE: *I'm still sitting where I was. I haven't changed my view. I'm still sitting at 49 and would reluctantly go down to 48. Every time I look at a 49, I think: "well that's a borderline script". If one's going to be kind, you could go to 48 but, I mean, anything lower... I'm sticking there.*

A3: *I've just looked at two 48s which are totally out of... We would be selling the pass I think if we accepted those, you know. I know its only two out of all the 48s but just looking at those two, if we said those were at A-level, then I think we would be selling the pass.*

A5: *Yeah, I've looked at two 49s which I would be, I think, happy to say had progressed significantly beyond GCSE and deserve a reward. And a 47 and a 48 which I would be very unhappy with.*

Chair: *Can I suggest that, for now, we put the boundary mark at 49 and then we'll have a look and see what effect this has* [for the examination as a whole] *when we've looked at Paper 3.*

The negotiating aspect of these interchanges is very evident. Although the awarders' judgements were presumably based upon relevant evaluative criteria (and the comments made by the same awarders during Step 1 indicated that, to some extent at least, they were) there was little attempt to discuss the evaluative reasons for the particular judgements reported to the group by each awarder during Step 2. Indeed, the one attempt to give such a reason by A1 was rather peremptorily, if probably correctly, dismissed by the Chair as inadequate. However, it would be a mistake to over-estimate the effect of this exchange

[*] SO = Subject Officer

upon the nature of the ensuing discussion which is typical of the observed meetings in general.

Awarder A3 began by suggesting a mark of 51 but, rather than giving qualitative reasons to support this judgement, used his second contribution to the discussion simply to prevent the mark from being set as low as 48. To achieve this, A3 appealed not to features of the candidates' work but, rather, to the other awarders' sense of collective responsibility for the maintenance of A-level standards in an abstract sense. Moreover, A2 and A4 both thought that 50 should be the boundary mark but had little impact because they expressed their views tentatively. Finally, the very definite recommendations for 49 from the Chief Examiner and A5 were accepted by the Chair despite his own repeated agreement with A2, A3 and A4 that 49 might be too low.

The one exception, of any extent, to the general failure to be explicit about the reasons for individual evaluations was in the English meeting. Here, reasons were more frequently, though still not routinely, given for judgements about scripts. (For example: *It's pretty close; yeah, there's just too many mechanical errors; there's too much..., it's too sort of slapdash and rather boring, really it's just telling the story* and *This one's competent; well, adequate but not inspired, the essay's not got much life but it's got some insight on Beckett.* and again, *It's got some good material and quite well structured but one essay's short.*) It seems reasonable to speculate that the use of such explicit evaluative comments may reflect the centrality of the evaluation of written material in English as a field of study. The English awarders may have been more comfortable with evaluative commentary by virtue of their normal professional work. However, even in English, reasoned evaluative discussion of the merits of individual scripts was a comparative rarity.

Given the general lack of specific evaluative comment, it was decided during the Phase 1 development work that the complexity of the initial coding scheme could be substantially reduced by removing the various subcategories relating to it. It was anticipated that analysis of the nature of the evaluative criteria used by the awarders could be based on the notes which they made during Step 1. It was decided that only the act of making an evaluative

comment about a script would be recorded during the meetings in Phases 2 and 3 of the observational work. Since the negotiation of the final decision characteristically involved awarders in suggesting marks, it was decided that the act of suggesting a specific boundary mark should be coded in Phases 2 and 3 and an appropriate category was introduced into the scheme to do this.

A further important point is illustrated by the above extract from the Physics meeting. The tentativeness of the suggestion made by A2 clearly reflects his difficulty in deciding on a single mark as the boundary. During Step 1, the remark from one of the Physics awarders that *This is so difficult* had met with general assent. Recognition of the difficulty of the judgemental task was common to all the awarding meetings observed. For example, in Mathematics the following, typical, remarks were made: *Well, I could live with 38. I saw two 38s and liked them both but there's a 39 I didn't like at all. And there's a couple of 36s and one of them, you know, well, it might just scrape through. I've only looked at one 37 and I... it wasn't quite there. But close, so 38 seems, you know, about right.* Some awarders accounted for data such as these by asserting that the original marking of some scripts was slightly lenient or severe and, of course, it is true that marking is never perfectly reliable. On the other hand, the awarders' judgements themselves are unlikely to be perfectly relable either, as most awarders freely acknowledge in informal discussion outside the meeting.

The board's officers in the meetings generally discouraged discussion of the reliability of the marking as being outside the awarding meeting's terms of reference. Since the awarders had not normally attended the meeting held to standardize examiners marking the scripts, this seems appropriate. Clearly, the awarding meeting must work with the mark scale as it was used by the markers if it is to make progress. A common strategy used by Subject Officers to inhibit discussion of the marking process was to ask for any particular script which caused concern to be passed to them for a further review by the Chief Examiner. Frequently, this review was conducted on the spot and sometimes the Chief Examiner would explain the marking of the script in question to the concerned awarder. Such exchanges sometimes produced agreement and sometimes produced agreement to differ but they

always served to establish the Chief Examiner, not the awarders, as the final arbiter of **marking** standards.

There is clearly a conflict between the awarders' recognition that their task is difficult and inherently imprecise and the procedural need for a single mark to be chosen as the grade boundary. More than this, however, there was an apparent conflict between the awarders' knowledge about the imprecision of their own judgements and their behaviour during the negotiations of Step 2 when considerable emotion can be generated in defence of a single mark. At first sight, this appears to be at odds with the results of Deutsch and Gerard (1955) that pressure to conformity with a group norm is greater if the judgement being made is a difficult one. However, awarders did not appear, in general, to treat differences between them of one or two marks as evidence of lack of agreement. Such differences were seen as being within the expected margin of error for the judgemental process or were explained away because all the awarders had studied overlapping, rather than identical, samples of scripts.

Thus, arguments in awarding meetings about differences in judgement of one or two marks did not, generally, reflect a conviction about the extreme precision of the judgements concerned. They reflected, in large part, the awarders' knowledge that their decisions were of vital importance to the candidates. Most awarders were evidently very concerned that their decisions should be fair to the candidates (see also Murphy *et al*, 1996). For example: *I really find this, you know, worrying. I mean I'm trying to be fair but here's a kid who's been let down by the teacher; he's bright but he just hasn't got the knowledge to use it. You can't say he's a B but you would if you could.* and: *I'd be really unhappy if this 76 didn't get an A, she's showing some real flair in places.* This latter quotation encapsulates the awarders' dilemma which leads to debate about single mark points during Step 2 of the Awarding process. Here was a script which the awarder believed to be worthy of a Grade A and so, in the interests of justice for a particular individual, he pressed for a boundary below 76, rather than 77. As one Communication Studies awarder put it: *I'd rather give a few candidates grades they don't deserve than penalise the ones we can't agree about.* In other words, the benefit of any doubt should be given to the candidates by awarding them the higher grade.

There is, of course, a counter argument, also based on notions of fairness to candidates, that unfair competition for jobs and higher education places will occur if some candidates receive higher grades than their attainment warrants. From this perspective, the grade boundary should be positioned so as to ensure that no candidate is given the benefit of any doubt about the grade which they should receive. The board's 1990/1 procedure file (Appendix 5.1) appeared to be taking a middle path: that errors of grading in either direction should be minimised. Addressing the imprecision of awarding judgements, It said:

> "In practice it is likely that awarders will agree that all scripts above a certain mark are worthy of a particular grade while all scripts below some other mark are not and that the provisional [ie. prior to ratification in Step 4] grade boundary for that grade is the mark mid-way between these two marks."

However, none of the observed meetings followed the approach described in this extract. All of them decided upon a single mark and arguments about justice to individual candidates, like those quoted above, were frequently voiced during the discussions. In the conflict between rather abstract considerations of fairness to candidates in general and the desire to give the benefit of any doubt to specific known (if only through their scripts) individuals, the latter proved the more powerful concern. This finding is consistent with the general literature on judgement and decision making. In particular, the power of individual concrete cases to overwhelm more general abstract considerations has been well documented by Nisbett *et al* (1982).

One of the arguments used, often by the Subject Officer, to defuse the emotional power of a particular script which seemed to be anomalous in deserving a different grade from those on the same, or the immediately adjacent marks, concerned the candidate's probable performance on the other paper. *If it's that good, he's probably done well enough on Paper 2 to get a Grade B in the end* is typical of the kind of remark made in these circumstances. As a strategy, this line of argument was generally successful, enabling the awarders to move on. Sometimes, however, it led into an instance of one of the more unexpected features of Step 2: discussion of awarding methodology. Reference has already been made to comments on the difficulty of the judgemental process. There was also occasional discussion of the types of scripts which facilitate or inhibit the formation of evaluative judgements. In general, awarders agreed that scripts which score very highly on some

questions, but very poorly on others, were much more difficult to evaluate than those which they call *balanced*. The following type of remark was very common: *It's no good looking at that one. It looks good at first but he totally messes up Question 4. You've really got to look at balanced scripts or you don't know where you are.* Methodological discussion appeared sufficiently common for it to be necessary to accomodate it by the introduction of a new category into the coding scheme.

Once qualitative judgements had been made at each key boundary for a particular component, most of the observed meetings then reviewed them from a statistical point of view. This was done by the Subject Officer using the component mark distribution to look up, at each key grade, the cumulative percentages of candidates awarded that grade or a better one. The following exchange from Step 2 concerning Paper 1 in the General Studies meeting is reasonably typical of the sort of discussion which followed.

SO: *Do you want to go over the percentages we've allowed through the various grade boundaries compared with last year?*
Chair: *[Indicates assent]*
SO: *Well, if I start with the A/Bs and work my way down. Last year we let 10.2% through at a mark of 40. This year with a mark of 41 we let 13.9% through.*
CE: *13.9 this year, 10.2 last year, yes?*
Chair: *So that is a move in the right direction. We were concerned we had too low Grade As last year so for that reason it is a move in the right direction and then, secondly, I think the examiners have indicated that the quality of the candidates was a little better this year.*
A1: *41 gives us a higher percentage and a higher mark so that moves us all in the right direction to some degree of statistical respectability.*
SO: *B/C is at 36 for both years. That's 22.6% last year and 28.7% this year.*
Chair: *That, again, seems reasonable.*
SO: *Now on the E/N boundary, 26 was the pass mark last year and that let 69.5 through and using 26 again this year would let 73.5 through.*
Chair: *Certainly don't want that any lower so if we stick with 26... OK?*
All: *[Indications of assent]*

It is clear from this extract that the statistical data at this stage were given little weight by the participants in the meeting (including the board's officer). The changes in the cumulative percentages of candidates at the three key boundaries were large, given that just over 1000 candidates were involved in both years. (The assessment of the statistical significance of such changes is addressed in detail in Chapter 6.) However, the awarders took the data in the way in which it was presented; as largely for information.

Only in one respect were the statistics of General Studies Paper 1 used to help decision making: there had been considerable dispute about the Grade E boundary earlier in the meeting and it had been left open at either 25 or 26. The Chair finally chose 26, apparently because this higher of the two marks nonetheless produced a sizeable increase in the pass rate. In general, this was the main use made of statistical data during Step 2: to resolve otherwise intractable arguments about the location of a boundary to the nearest mark. Indeed, Subject Officers in many of the meetings suggested leaving the final precise placement of component boundaries at the centre of such disputes until the statistical data were considered towards the end of Step 2. This tactic gave time for any emotion which might be blocking agreement to dissipate as well as introducing extra information to help resolve the deadlock arising from the qualitative judgements alone.

One other feature of the above extract from the General Studies meeting is worthy of comment. The examiners' views, expressed during Step 0, that the current year's candidates had higher attainment than last year's were quoted as accounting for the rise in candidates' grades. Logically, this is extremely questionable since those same examiners had contributed to the judgements made at the meeting and their prior judgements were the result of the same evaluative process as was used in the meetings. They could hardly, therefore, provide independent corroboration of the results of the meeting. Nonetheless, the use of the examiners' initial reports to justify the statistical consequences of the meeting was common to most of the awarding meetings observed in Phases 1 and 2. Formally, of course, it is only as a result of the meeting itself that it can be possible to say whether one year's candidates are better than another's. Indeed, that is the very purpose of awarding meetings.

5.5.4 Step 3 - combination of the component decisions to produce a boundary for the examination as a whole

At the time of Phases 1 and 2 of the study, the board's procedure (Appendix 5.1) was to use only the addition method (Model 1 from Chapter 4) to combine the component boundaries to produce subject boundaries. This process was carried out by the Subject Officer in each meeting. Occasionally, one or more of the awarders carried out the calculations in parallel; more often, the Subject Officer did them alone. Clearly, the calculations were more

straightforward for those examinations in which no scaling of component marks was required.

Some Subject Officers followed the practice of combining the component boundaries at each key grade as soon as a decision had been made for the grade boundary on the last component to be considered. Others, however, waited until all three boundary decisions had been made on every component before beginning the job of combining them. Since the awarders were generally keen to hear the overall consequences of their judgements, which would determine the course of Step 4, Subject Officers following the second of these two approaches carried out their calculations under rather more pressure than their colleagues adopting the other approach. More importantly, Subject Officers who combined the component boundaries as they became available had time to reflect privately upon the outcomes while the awarders were carrying out the scrutiny of scripts for the last component boundary. This gave them a better chance of effectively managing the discussion during Step 4 of the awarding process.

Once Subject Officers had computed the subject grade boundaries, they used the subject mark distribution to look up, at each key grade, the cumulative percentages of candidates awarded that grade or a better one. These cumulative percentages were the main data considered during Step 4.

5.5.5 Step 4 - ratification of the examination boundary

In all the meetings observed in Phases 1 and 2, Step 4 began by the Subject Officer reading out the cumulative percentages of candidates at each key subject boundary. There followed a discussion of these data. If there was little change in the cumulative percentages from those of the previous year, Step 4 was brief. All participants expressed themselves satisfied with the outcomes, taking the statistical stability as evidence that the qualitative judgements were satisfactory.

During Step 4 a clear change in the involvement of the Subject Officer was observed in most of the Phase 1 and Phase 2 meetings. In the earlier stages, the concerns of each meeting centred on the formation of evaluative judgements and the Subject Officer had an almost exclusively administrative role but, during Step 4, the views of the board's officers became more significant. In particular, it was the Subject Officer who usually decided whether a "marked" (see Appendix 5.1) change had occurred in any of the cumulative percentages. If he or she thought that such a change had occurred then he or she would either suggest to the awarders that they should reconsider particular judgements or would call in a member of the board's senior or research staff to discuss the matter further.

An example of this procedure working apparently effectively to deal with anomalous statistical data uncovered during Step 4 was provided by the Mathematics meeting observed during Phase 2 of the fieldwork:

SO: *Grade A had 20.4% [of candidates the year before] as against 16% this time, there were 37.2% as against 29.8% at Grade B, so we're 8 lower there. And 77.1% and 67% for E, so the results are a lot worse this year.*

CE: *Well, Chris [a senior examiner] said it was a far worse performance everywhere this year than anything she'd seen before in her life. You know, sort of, astonishingly weak. But it looks as if the performance in Pure Mathematics in general is even worse than she... I do think the standard is down.*

A1: *It may not be the pupils. It may be the teachers and the effects on the teachers of what's been going on in schools... I mean, you know, it may not be the staff. It may be what's happening in the school system that's causing this.*

A2: *But its very unusual on an examination paper to require the pass mark to be at the mean and that would be something that we, one, would find very hard to defend. Have we done it before?*

SO: *I think we've got problems. What we've normally done is compensate with another paper. We certainly did it on Paper 1 one year when the statistics on Paper 2 shifted. Of course, when the statistics go sort of in candidates' favour, not many people complain but we've shifted the statistics against the candidates from last year and if we stick with that, we will get, I am sure, people asking why the candidates have not done so well as last year. We had two lowish means once before and we never compensated enough for those low means and we had a fair number of re-marks and letters saying the centres did not find the results satisfactory.*

A2: *We're asking for 38 on a paper with a mean of 61.4 [Paper 1] and for 34 on a paper with a mean of 34 [Paper 2]. The only way we're going to make sense of this is to... actually to drop that and maybe go up on Paper 1 to spread it a bit more.*

A3: *I think the way you could justify these things is if there was a substantial proportion below 10 - much more than last year. I mean, if there was real*

evidence that there was considerable weakness. I know there can be harder papers with lower marks but there could be lots of very weak candidates on Paper 2.

A2: *Who aren't weak on Paper 1?*

A4: *Well, talking personally, we [the awarder's own college] have a fairly consistent group of students nowadays and they certainly found it very, very difficult this year. And the teacher who taught them... I mean, she was an experienced teacher but they came out... I mean, I was quite astonished when she passed me the paper and said, you know, "Were you thinking of awarding these things? Were you at the moderation meeting for that? How does such a difficult paper get approved?" And so I anticipated it was going to be a problem because I had thought it was, it seemed to me, a well-moderated paper. but what has been said about it is obviously true and yet I don't understand why we didn't see it. I personally always try to make sure that they can get into the questions and yet they didn't get in and when you look at the scripts that was the problem. They floundered at the beginning of the questions really. Even the strong ones floundered and yet they performed in the sixties on Paper 1. It has to come back to the paper.*

CE: *Well I don't think it is hard if you have taught the work. I don't think it's more difficult, its just more maths.*

A2: *Yes but isn't there an intrinsic perception that this paper is for double mathematicians. Doesn't that affect the way you look at it. I mean, you know, this paper is nice - it's an elegant paper - but doesn't it require much more... insight than the other papers?*

A4: *I think that's right.*

Chair: *So, how do we proceed?*

A3: *Can we know the length of the tail? [of the Paper 2 mark distribution]*

SO: *Last year on Paper 2, the mean was 54.7 and the standard deviation about 30, we did have 31 candidates with zero, if that means anything. And this year we're up to 85 on zero and then going upwards we have more candidates at every mark between 0 and 10. Making a very quick comparison we've got 30s and 40s as against 15s and 20s.*

Chair: *So, for example, the number of candidates with 10 marks or less? do we know?*

SO: *Well, on 10 marks this year its 84.5% cumulative [from maximum marks], last year it was 90.9 so there's always about 10%... well, not always but on the historical evidence of one year, there does appear to have been a tail last year as well.*

Chair: *But a more substantial one this year.*

A2: *There's always a group of people who try hopefully.*

Chair: *Any further information anyone would like to try to elicit please?*

A5: *It's the relative largeness of the standard deviation against the mean that worries me in terms of what sort of distribution has been thrown up - whether the distribution is anything like normal or abnormal.*

SO: *It looks amazingly flat right the way across. The number of candidates is about 40 at a mark of about 50 and then nearly every mark below that we've got somewhere between 35 and 50 candidates.*

A5: *It's not aberrant. Therefore there'd be more justification in actually... if you do move the marks down it's fair because you don't get an odd bump in the distribution.*

Chair: *No, you've got almost a uniform distribution.*

A5: *Is there the same sort of numbers... well obviously there isn't the same sort of number at the top of the range is there?*

SO: *No, the top of the range is very thin of numbers. We're talking about very few candidates. There are no marks above 70 with candidates in double figures.*

A5: *So, if you dropped the A/B boundary it would have some consistency.*

Chair: *Yes, if you dropped the A boundary you wouldn't suddenly have an enormous...*

A5: *That's what I was worried about.*

A4: *Would it be a similar sort of thing at the other boundaries, if they're sort of even too?*

SO: *It would. B/C would be more or less the same. B/C was 59. It's fairly flat.*

A2: *I don't think we can justify only half the candidates passing [on Paper 2].*

Chair: *No, I'm sure.*

A2: *We just can't justify that.*

SO: *I mean, the horrifying thing is that if you wanted the same number of people to pass on Paper 2 this year as last year, you'd have to come down to...*

A5: *you'd want it like '89, rather than last year, I think.*

SO: *Yes, but you'd have to come down to a mark that I don't think anyone would be prepared to accept in any way.*

Chair: *OK, in '89, Paper 2 had a mean of 49.7 and we took the E/N boundary at 39. So we're talking... in '89 we took about 10 marks below the mean. Well 10 marks below this mean, you can see for yourself - 24.*

A5: *You have to take it as a proportion of the marks, rather than an actual number. You're starting about 10 further down so you've got to take about 8 less.*

Chair: *I've got a nasty feeling that we're going to be plucking figures out of the air at this point.*

CE: *Is there any relevance of looking at the same scripts of some of the people on the boundaries on Paper 1 and looking at their scripts on Paper 2. In other words, find the same candidates who are on the boundary. I mean, it might take some time to do it but I think it's far more fair to have a look at the combination of the two scripts and make the judgements on the basis of that, than on plucking figures out of the air, because you are looking at their whole performance.*

SO: *So, you'd like us to find...*

CE: *I know it may be asking a lot John [the Subject Officer].*

SO: *Well, we could find... well to some extent.*

A2: *Do you have any provision to make grade awards on percentage... on percentiles, rather than... I know it's not normally the practice but it is a practice to make an award on percentiles, rather than on marks.*

SO: *The simple answer to that is: not specifically. The awarding committee can, with the judgement of scripts and the basis of statistical information, make whatever final decision they think is right.*

CE: *Do you have on the computer the scores of each of the candidates on the individual elements?*

SO: *Yes, I have them here. [as hard copy]*

CE: *Well the other way would be to have a look at some of the borderline candidates we've talked about and just call out their scores on Paper 2 just to have a look and see what they've got.*

Chair: *That seems a very good idea. I favour that, rather than talking about percentiles or any other method.*

CE: *Well, we'd get some idea whether that 34's got any meaning or not.*

SO: *In which case... I've got 38 for Grade E on Paper 1. We've got 28 candidates who got 38 on Paper 1. So if I read out what they got on*

Paper 2. I doubt I'll find all 28 but if I get 24 or 25 of them I'm sure you'll be satisfied. First one is 6.

A2: *Bloody Hell! [pause] Is this being recorded?*

All: *[laughter]*

SO: *22, 20, 18, 17, 19, 14, 19, 12, 18, 18, 15, 5, 22, 9, 14, 13,*

A5: *I don't think you need to do any more. The plain straightforward thing is: there's not a single one there half a standard deviation below the mean on that paper.*

A2: *Can't the boffins help you?*

Chair: *Well we are entitled to call in expert assistance.*

A4: *What worries me is that if there was, amongst these scores you read out, a few marks around the 30s, I would be fairly happy. But there's not a single one above 22.*

SO: *We've got to the situation where we are asked to maintain our standard from year to year - and if our standard in '89 was right and if our standard in '90 was right, given roughly the same number of candidates, certainly as 1990, but fewer than 1989, we are about to award grades that are less than the previous two years.*

A4: *Significantly*

SO: *Yes, very significantly less than last year and significantly, well 5% or so, on 89. And so the Committee is really asked to consider: the judgement of scripts takes us - we've got as low a mark as possible - whether it wishes to move the subject boundaries in some way to give a situation closer to previous years.*

A2: *Can the Committee look at percentiles as I asked before? Or else can the research department construct a linear model for us which takes into account the difficulty index of the two papers, which is a statistical technique which can be used.*

SO: *All you have to do is just look at the subject mark distribution and say what mark has to be obtained... what mark would we have to come down to. Is the Chief Examiner, are the Committee, prepared to do that?*

Chair: *Again, well, my gut feeling is that if we come down as far below the mean as we did in 1989... I was hoping we might find a few figures around the 24 mark but the nearest we've got to it is 22.*

CE: *That's right. This is what worries me because I was hoping that we might see what evidence we could pick up in the higher 20s, lower 30s. My guess would be that if you... because we were trying to see whether there was a figure which, if you go below 34, would make some sense. Well there isn't.*

SO: *What we are facing is... taking those candidates who've got 38, having taken Paper 1, had they taken Paper 3 [which, with Paper 1 forms a different option within the Mathematics syllabus] may well have got 45, having taken Paper 2 they've been torn apart. I mean, the Paper 2 mark, you wouldn't be surprised to hear, if you wanted to go to roughly the same percentage as last year... you'd have to come down to 22. Which is why none of those borderline candidates have got above 22.*

A5: *Normally speaking our bottom mark is around half a standard deviation below the mean.*

A4: *But this is a common problem with papers which have been found to be difficult in the exam. The culture shock of actually going so low means that people never get as low as they should do. They always finish up with a compromise. You know, 27 might be the compromise when, in fact, you want to go to 22. The problem at the end of the day is that the candidates are suffering with the figures we were suggesting before.*

A2: *But the other board that operates statistically is quite happy to go down to quite low figures. I remember 19, once.*

Chair: *Well ladies and gentlemen we still have to come to a decision.*
SO: *Can we look at those figures which would bring us comparable to '89 on the subject distribution and then see what that would involve?*
A5: *Yes, the whole distribution looks nearer to '89 than '90 anyway.*
SO: *Because that way we would have some historic... what we're doing now is saying that the paper hasn't worked as well as we would have liked and we're making a statistical historical judgement on the standards. I think we needn't go up to the figures of 1990 because they were rather higher than some.*
CE: *Well, all the evidence is that this year's are getting more like '89 anyway.*
SO: *OK. We had 72.1% passing in 1989. Coming to 72% would mean 65 marks for the subject so we would have to come down as far as 27 for Paper 2 this year.*
Chair: *Well it's not so bad, not as much, as I thought.*
SO: *But it's below where the judgements were made.*
CE: *Yes, but if you actually looked at candidates who've got that total, you're probably going to find that they got maybe 40s [on Paper 1] and maybe teens [on Paper 2].*
SO: *Yes, you might find that was what a balanced candidate got. Anyway, that would be the figure. At A/B to get 20.4%... at 20.2% you get 140 marks so we'd have to come down 7 on Paper 2 to 63. We'll leave the B/C for a moment. I haven't had a chance to check if '89 is comparable with '88 and '87. '87, I think was a hardish year.*
Chair: *In fact, in '89 we gave more Grade As than in any of the preceding years.*
SO: *So, historically, the justification for heading towards the '89 figures is?*
A1: *What you're saying is we've got to knock 7 off the E/N total but you're obviously not wanting to knock so many of the A/B total. Remember that somebody did actually suggest a mark of 67 as being a grade A [during Step 2] although we finally picked a figure of 70 so that there is some room for manouvre at the top end.*
SO: *We've got to be careful about getting the right passes. I mean, if you look historically, in some ways the comparison should be with last year but maybe we think last year might have been over generous?*
A2: *Last year* ***was*** *a good year.*
SO *Yes, well in '85, '86 and '87 the pass rate was only about 63%, then '88 was 67%, then up to 72%. Then we were up in the 70s and last year was 77%.*
Chair: *We have been trying to bring it up, haven't we?*
SO: *Well 27 and 63 are the marks we need to bring it back to '89's figures. If we compromised somewhere within there and chose 67 [for Grade A] which was actually mentioned and if we compromised on 30 [for Grade E]?*
A2: *I think, given these marks that we've just seen here for the Paper 1 borderline ones, 30's not going to do any of them any good is it?*
A4: *I agree. I'd like to see that 27 stay as 27 or even go down lower.*
SO: *Well be careful, because we've got nearly 80% with 38 or more on Paper 1. We know that in the past candidates have got their Grade As by getting 90s on Paper 1 and actually getting a B or a C on Paper 2 but getting enough marks to pull them up.*
A2: *I think you've also got to accept the fact that we were fairly happy with the outcome of the judgements on Paper 1. There's no need to change that because of an aberrant paper. The question is, now, do you in fact... all you can do is make changes on the paper that's different in order to get some sort of justice into the system. And I think, I agree with [A4] here, I think we can't compromise on that 27 figure for the pass mark. I think you've got to go for 27.*

Chair: I agree. I think you've got to go certainly that far.
CE: What percentage, then, get 65 did you say?
SO: 72%
CE: That's not so bad. That sounds better for people doing double maths.
A3: From what's been said it sounds as though the best you can do is go to 27 for the E/N boundary, 67 for the A/B boundary and somewhere in the middle for B/C.
SO: If we go to 67, we will have 18.2% Grade As. and B/C would be about 56/57?
A2: 56
SO: 56 gives 122 for the subject - 32.1%
A2: I think that's justifiable
SO: And E/N for the subject was 65 that's...I've already said that, its 72%. How do those three figures match with '89?
A2: The E/N matches accurately
SO: Yes, it must, but the others are still 3 or 4% down.
Chair: That doesn't worry me so much.
SO: 18% is 2% down on '89 and the other one is 5% down.
A1: Well last year's papers were good papers and the candidates performed well but this year the performances were worse.
Chair: Well, those are the figures we've now come up with. Does anyone want to make any further comments or adjustments?
[Further discussion focussed upon the need to review the following year's papers to avoid a similar situation recurring and eventually the statistically-derived grade boundaries were agreed by the group.]

Several interesting features of this lengthy extract deserve comment. First, there had clearly been a failure on the part of all concerned, examiners, committee members and Subject Officer, to appreciate beforehand how difficult Paper 2 would prove to candidates. There are hints of one possible cause of this in the remark about the elegance of the paper; it seems possible that the paper's aesthetic appeal to experienced mathematicians blinded them to the difficulties it posed to candidates. Whatever the cause in this case, the unpredictable nature of the difficulty of assessment instruments in general and the implications of this for the assessment process were discussed in Chapter 3. Some of these implications are graphically illustrated by the events in the Mathematics award observed during Phase 2 of the present study. In particular, the awarders' original qualitative judgements led to surprisingly large reductions in the pass rates for each grade they considered. This is consistent with the results of Good and Cresswell's (1988) study which clearly established a tendency for awarders to make relatively severe qualitative judgements on relatively difficult papers and/or relatively lenient qualitative judgements on relatively easy papers. This phenomenon is discussed in more detail in Chapter 6, where data from a range of

examinations are presented which imply that it tends to occur even for relatively small variations in the difficulty of the papers being awarded.

In the Mathematics award, the results of the original qualitative judgements of candidates' scripts in Steps 1 and 2 were evidently a major concern to the Subject Officer and several of the awarders. Despite a few initial attempts to explain the results away, the meeting rapidly came to see them as problematic. This is the second point arising from the extract from the meeting. There was evident concern that, for reasons of fairness, the award should not be more severe than those of previous years and an acceptance that the large changes in the statistics of the results indicated that it would be more severe if the qualitative judgements were accepted.

The Mathematics awarders therefore began to cast around for a procedure, other than the simple acceptance of their own original qualitative judgements, which would enable them to be more lenient. This was necessary because, during Phases 1 and 2, the board's procedures did not specify how the statistics should be used beyond the *identification* of a problem of comparability. (The possibility of taking technical advice was raised repeatedly by Awarder A2 but was not followed up by the Subject Officer. It may be that the presence of the author, who would normally be available to give such advice, observing but not participating in the meeting had some effect on this aspect of the meeting's procedings.) Eventually, the awarders made two key decisions: to make the year of comparison the year before the previous year and to rely almost exclusively upon the statistical data, rather than qualitative judgement of the scripts.

The justification for the first of these decisions was essentially pragmatic and reflected the awarders' unease at adopting a purely statistical approach: they did not want to use an extreme set of results as their benchmark. The second decision effectively acknowledged that the qualitative judgements were invalid in the circumstances of an unexpectedly very difficult paper. Considerable time was spent achieving a consensus on this matter with Awarders A2 and A4, together with the Chair and Subject Officer, seeking to carry Awarders A1 and A3 and the Chief Examiner with them. One clear feature of the arguments used to

do this (by Awarder A2, in particular) was the effort made, with apparent eventual success, to legitimise a very low percentage mark as the pass mark. As noted in Section 5.5.2, most awarders appear to feel intuitively that, in terms of percentages of the available marks, grade boundaries should not vary much from year to year. The relative rarity of major changes in difficulty, such as that exemplified in the Phase 2 Mathematics award, probably contributes to this widespread intuition since, even from the *script as response* perspective (see Chapter 3), it is correct, provided that the difficulty of the paper concerned does not vary much from year to year.

However, the relative readiness with which the qualitative judgements were set aside in the Mathematics meeting was atypical. In all the other observed meetings, when apparently large changes of statistical outcomes occurred the awarders were reluctant to amend their original judgements in the light of the statistical evidence. The following extract from the General Studies meeting illustrates the sort of debate which more typically occurred.

SO: *Well, when you crunch the numbers together, then the overall boundary mark of last year was 277 which was the lowest mark in A and this year it is 279. Last year we had 1.8% got through at Grade A, this year we have got 5%. For B, 248 was the mark last year and the mark this year is 246 - very close - and 18.3% got through this year whereas only 11.2% got through last year. And finally, the mark at the E boundary was 170 last year and it is 168 this year and last year 75.9% got through; 85.4% got through this year.*

All: *[Reactions of surprise]*

Chair: *The Chief and Senior Examiners testified to the rising standards.*

SO: *85.4%, a 10% increase.*

A1: *Reducing standards during the day?...*

A2: *Every year!*

All: *[laughter]*

SO: *It looks rather generous.*

Chair: *We have got to go to the Committee with Computer Studies breathing down our neck asking to have it explained away [a reference to the relevant subject committee of the Board which was also responsible for Computer Studies].*

CE: *I don't feel like explaining it away.*

Chair: *E for Paper 1 was 26, for Paper 2... 40 and for Paper 3 was 34. Now if I recall, I cannot remember as far as Paper 1 but certainly for Paper 3, [the Chief Examiner] was hesitating on 34/35. I don't think you can really do a lot of movement there. I think, in support of the movement towards an increase in coursework percentages, he said they were beginning to understand the type of assignment. I don't think there was much argument over Paper 2.*

A1: *Last year was the first year of the presentation. That's what we said.*

SO: *You had 69.5% for last year on Paper 1 and you've got 73.5% this year.*

A2: *Four points.*
SO: *And on Paper 3 it has gone from 80% to 87% so that is a bit of a jump.*
Chair: *It has gone up 7.5% compared with...*
SO: *Paper 2, of course, has gone up too. It has gone up from 69.5% to 73.5%.*
CE: *I am happy to run those percentages myself. Do you think the Board will not be happy with that?*
SO: *I don't think they will be unhappy about the higher grades because they were somewhat low and out of line with other subjects.*
Chair: *OK, but when you have got the average pass rate for A-level around 75% and we have been running with that for many a year now, and now all of a sudden we are jumping 10%. I just don't believe the students are that much better.*
SO: *I think I should consult someone.*

[A more senior member of the Board's staff (SS) was asked to join the meeting at this point and was told of the statistical outcomes of the current and previous years.)

SS: *I would be happy with the rise in percentages if I knew why the [boundary] marks have gone down. The percentages might be because of a better entry and the A boundary has gone up by 2 but the B and E boundaries have gone down by 2 even though the mean mark is up this year.*
CE: *Yes, but we've been concerned about the low success rate. I think we felt before that people were being put in without being properly taught or prepared for the course. There was quite a lot of evidence for that. Last year, maybe people were being put in prematurely.*
SS: *When you made the decisions this year, were you consciously carrying forward last year's standard or seeking to modify it in some way?*
Chair: *No, we were concerned about Grade A but we didn't drop the standards to get a higher pass rate.*
A2: *I think, at the outset, what was commented on was that the standard was better. Before we started the process we went through the papers. The candidates had been better prepared and were better able to answer the questions.*
CE: *It is certainly interesting, I went through this beforehand with [Subject Officer], that the improvement in the mean marks was consistent throughout all three components. That is actually fairly unusual to get that consistency through all components. Just a small rise in each in terms of performance. The message coming through is that there has been a change in performance more than any easing of the requirements for a paper.*
SS: *There does seem to be a good deal of consistency in the way the figures have come out. In fact, all the marks you have put in are close to last year's. I think the grade boundaries [on the components] have not changed by more than 1 mark at any point. So it is all very tight and consistent and you think this year's candidates were better prepared?*
CE: *Yes.*
SS: *Well I don't see why the Board should worry too much about this. We've looked at the statistics. They're a bit surprising but we can explain them and the top grades have probably always been a bit low in this subject. I think you should go ahead with what you've got.*
Chair: *As long as the Board is happy... There will be a lot of happy people in the schools.*

Several significant points are illustrated by this exchange. First, the surprise expressed by the awarders when the Subject Officer announced the results is itself surprising. At the conclusion of Step 2 for each component, the awarders had been told of the statistical consequences of their qualitative judgements. These had all shown increases in the proportions of candidates with marks exceeding the grade boundaries. Evidently, the awarders had not appreciated the extent to which the component changes would combine to give more candidates passing each grade boundary for the subject as a whole.

The second point to note is the determination of the Chief Examiner not to modify the judgements made on the basis of the scrutiny of scripts in the light of their statistical consequences. This was a very common attitude during Step 4 of the Phase 1 and Phase 2 meetings and the approach of the General Studies meeting was more open to change than many. Indeed, comments like the following are common: *I can't see what the point of looking at the work is, if you want us to change our minds because we don't agree with the statistics.*

It is asking a lot of people who have spent most of a day (or longer) making careful qualitative judgements that they should reconsider those judgements because of information which was available all along but not revealed to them. The rationale for the board's procedures in this respect was to avoid prejudicing the qualitative judgements with the statistical evidence. However, the effect was to build conflict into the final stage of the awarding process and, effectively, to limit significantly the extent to which the statistical data could influence the decisions being made. It may be that the low weight given to the statistics in the awarding process as a whole was a matter of deliberate policy but the observed behaviour of the board's officers during Step 4 was inconsistent with the existence of such a policy. The officers clearly took the statistical data seriously and were concerned if the outcomes differed much from one year to the next.

However, some awarders refused to accept that anomalous statistical data were relevant and there was a tendency to explain them away by maintaining that, as a group, the candidates were simply better (or worse) this year. This assertion was sometimes

supported by reference to the balance of centre types and/or genders of the current year's candidates or to the effects of an increasing or decreasing entry. These sorts of explanations are evaluated in detail in Chapter 6. Even when the awarders accepted the relevance of the statistical data, they tended to see any anomaly with the previous year as a problem for the board and its officers, rather than for themselves. Since the awarders had the final decision-making power, it was possible for them to ignore anomalous statistical data in Step 4 and leave any problem it caused to others to solve. Some awarding meetings clearly took this course and this seemed to occur particularly when the status normally enjoyed by the Subject Officer *vis-vis* the awarders was not high. In connection with this, it is worth noting that sometimes Subject Officers' entirely administrative role in the earlier steps of the meeting appeared to undermine their professional status *vis a vis* the awarders and make it easier for their arguments to be discounted during Step 4.

The awarders often appeared to feel threatened if it emerged during Step 4 that their judgements were producing large changes in the statistical outcomes of the examination and one of the most common defensive reactions was for the awarders to define themselves as a social group excluding the Subject Officer and other staff of the board. This was particularly evident if another member of staff was asked to join the meeting during Step 4 to discuss the statistics. *Here's the statistician, come to sort us out.* was a typical reaction in these circumstances. The well-established tendencies of members of groups in conflict to devalue information coming from members of other social groups, to perceive members of other groups in stereotypic ways and to become more certain of the accuracy of their own collective judgement (Baron *et al*, 1992) were clearly present. This is illustrated particularly well by a vignette from one of the Phase 1 meetings where the Step 4 debate had become particularly intense. One of the board's Research Officers had been asked to join the meeting and had been attempting to persuade the awarders to reconsider their judgements at the Grade A boundary. One of his arguments was greeted by the remark from an awarder that he was *challenging our professional judgement* as if that ended the matter. In response, the Research Officer referred, apparently quite deliberately, to his own professional status: *I am simply saying that, in* ***my*** *professional judgement, these statistical changes are very hard to justify for such a large group of candidates from much the same*

schools when the same changes haven't, apparently, happened at the other grades. This was greeted in quick succession by reactions of surprise and then a more constructive debate as the implication of the officer's remark, that his conflicting view should be given appropriate status, was made explicit by the Chair.

These sorts of effects were evident even in the more relaxed atmosphere of the General Studies meeting. The extract above shows how, when the meeting was joined by a member of the board's senior staff, the awarders clearly felt that their work was being evaluated as, in a sense, it was. The meeting took on a more formal tone and the Chair responded very quickly to rebut any suggestion that the awarders had deliberately reduced their standards. The discussion took the form of the board's senior officer raising questions about the awarders' judgements and the awarders justifying their conclusions. Such an approach is likely to produce a tendency on the part of the awarders to stick to their original judgements. For example, the General Studies Chair did not reveal to the member of senior staff his concern, clearly expressed earlier, that the Grade E boundary may have been too lenient.

In general, the apparent task of the board's officers during Step 4, to bring the awarders to a serious consideration of the statistical evidence, proved a very difficult one. It is not, therefore, surprising that the statistical outcomes discussed in Step 4 had little influence on the final positions of the grade boundaries in most of the Phase 1 and Phase 2 meetings. Justifications for using the available statistical data in this way and the question of the weight which can reasonably be given to them are considered in detail in Chapter 6.

The final point illustrated by the General Studies and Mathematics extracts given in this section is the tendency for awarders and Subject Officers more easily to accept an increase in the grades awarded than they do a similar decrease. Steadily improving grades can be interpreted as an indication that the syllabus is having desirable educational effects and, from a narrow administrative perspective, is likely to reduce the number of candidates appealing against their results. There is therefore little incentive for awarders to question improving statistical outcomes particularly when they follow from the awarders' own professional judgements, formed over several hours of hard work.

Most significantly, it was established in Section 5.5.3 that awarders generally subscribe to a principle of giving the benefit of any doubt to the current year's candidates even if, in so doing, they may risk failing to set standards comparable to the previous year and, thereby, unfairly damage previous candidates' selection chances. There was a general tendency, evident in the remarks made during awarding meetings, to regard unexpectedly easy papers as a success since they were held to enable the candidates to show what they know, understand and can do (for example: *The paper worked well this year, the kids could really show what they could do*). There was then a reluctance, driven by concerns about fairness, to discount some of the candidates' performances in the light of the easy context in which they were produced. (The argument was that any attainment that had been demonstrated must be rewarded; but this is implicitly to adopt the *Script as Artefact* strategy, a strongly criterion-referenced approach to awarding, the naiveté, and inappropriateness of which was discussed fully in Chapter 3.) The conditions were thus entirely right for group polarization (in the form known as the *risky shift*) to occur. This is the tendency for groups acting collectively to take greater risks than they would as individuals and it occurs particularly easily when, as individuals, the members of the group already share a willingness to take the particular action in question (Brown, 1988). On the other hand, as the Mathematics meeting illustrates, the importance of statistical comparability with previous years was more readily accepted if the current year's qualitative judgements looked likely to be more severe and therefore to risk being unfair to the current years' candidates.

As far as the observational coding scheme was concerned, there were no aspects of the discussion in Step 4 of the awarding procedures which did not arise during earlier stages. No modifications beyond those already described were therefore necessary, during coding scheme development in Phase 1, to accommodate the Step 4 discussion.

5.6 THE PHASE 3 OBSERVATIONS

5.6.1 The changes in procedure introduced prior to Phase 3

Appendix 5.2 contains a copy of the procedure paper used in the Phase 3 awarding meetings. Comparison with Appendix 5.1 reveals a number of significant differences which are summarised below:

1 New instructions were given to Chief Examiners about their preliminary reports. These emphasised that the reports should focus upon the demands of the current year's papers compared with those of the previous year and not include speculation about the general level of attainment of the current year's candidates.

2 The use of statistical data to select the scripts for scrutiny in Step 1 was introduced. From 1992 onwards, the scripts scrutinised by the awarders for each grade boundary, on each component, were selected to cover a range of marks either side of the mark at which the cumulative percentage of candidates was as nearly as possible the same as for the same boundary in the previous year.

3 Statistical data showing how the mark distributions for the current year compared with those of the previous one were given to the awarders by the Subject Officer at the start of the meeting and before the Step 1 consideration of each component.

4 The awarders' instructions were modified to make it clear that the grade boundaries they adopted should be based on both the evidence of their own evaluations of candidates' scripts and a consideration of the above statistical evidence.

5 For medium and large entry examinations, expected changes in subject grade distributions were established. Awarding Committees wishing to adopt grade boundaries which produced changes greater than these expectations were required to provide a supporting written rationale.

6 The role of the awarding meetings was changed from decision-making to making recommendations about the location of the grade boundaries. Responsibility for the final decisions was transferred to the Secretary General, representing the Board itself.

7 As well as being valuable in their own right, these changes were also intended to enhance the status and influence of Subject Officers vis-a-vis the awarders by making them the principal interpreters of the statistical data mentioned in Points 3, 4 and 5 and, most significantly, the people who presented the meeting's recommendations to the Secretary General.

8 The procedure recommended in SRAC (1990) was adopted for combining component boundary judgements in Step 3. That is, for each key grade either the mark produced by the addition method or the mark produced by the percentile method, whichever was the lower, was used as the aggregate boundary.

These new procedures were introduced for the Summer 1992 awarding meetings and, prior to their introduction, a one day briefing was held for all chairs of awarding committees to explain the new procedures and their rationale. A briefing paper was issued to all awarders explaining the new procedures and making the case for greater use of statistical evidence. Training courses were held for Board officers on the new procedures and the management of awarding meetings.

5.6.2 The observable effects of the new procedures upon Steps 0, and 1

The new procedures involved a greater formal use of statistical data in Steps 0 and 1 of the awarding process (Points 1 to 3, above). The main consequence of this which was observed in Phase 3 of the study was the operation of the new procedures themselves: the scripts scrutinised in Step 1 were no longer chosen on the basis of the Chief Examiner's qualitative judgement and the awarders were given information about the statistical

equivalence of the marks in the current year and those in the previous year, before they scrutinised the scripts.

However, the nature of the Chief Examiners' reports in Step 0 changed little, despite the emphasis in their briefing on the need to discuss the papers, rather than the candidates. Indeed, one Chief Examiner explicitly expressed doubt about the value of describing the functioning of the paper to the rest of the awarders, taking the view that this was completely predictable. Another Chief Examiner explicitly disagreed with the implications of the statistical data for the boundary marks: *Whatever the statistics might show, I don't think the paper was any harder this year and I would recommend keeping the boundaries in about the same place as last year.* Of course, it is not unusual for people to resist changes in working practice which are imposed upon them and reluctance to adopt the new procedures could have been transitional effects. However, the Phase 3 observations were made in the second year (1993) of the new procedures and later informal observations of meetings in 1994 and 1995 confirm that more needs to be done if Chief Examiners are to be persuaded not to assume that changes in the level of marks awarded on a particular year's paper are **necessarily** due to improved or worse performance on the part of the candidates.

The new procedures also had little apparent effect on the behaviour of the awarders during Step 1, but this is neither surprising nor a cause for concern. In the informal discussion of scripts which is a hallmark of Step 1, awarders continued to refer to their own professional judgements of the quality of the candidates' work and increased references to the statistical evidence were not apparent. On the other hand, the awarders did not find the scripts provided (on the new statistical basis) inappropriate and, in every case, judged some of the scripts to be one side of the boundary being considered and some to be the other. In none of the Phase 3 meetings did the awarders ask, as they could have done, to see scripts outside the range of marks represented by those initially provided.

5.6.3 The observable effects of the new procedures upon Step 2

The new procedures did have a clear impact upon the discussion in Step 2 during some of the awarding meetings observed in Phase 3. The essential negotiating nature of the discussion was not changed but significant references to statistical evidence were evident in all the meetings. In this respect, the approach taken by the Chair of each meeting was crucial. The special briefing meeting held for chairs appeared, in most cases, to have changed their perception of the process and their role within it. They tended to present themselves less as one of the awarders, and more as an arbiter between the examiners, who continued to emphasise qualitative judgement, and the board's officers who drew attention to the statistical data. As a result, increased participation by the board's officers at this stage in the meetings was evident and their statistical contributions were treated as more relevant to the task in hand. All the observed meetings seemed to attempt to make a composite judgement at this stage on the basis both of their qualitative judgements and the statistical data presented to them by the Subject Officer. These changes in the behaviour of the meetings' participants in Step 2 are confirmed by the analyses of the systematic observations reported in Chapter 6.

5.6.4 The observable effects of the new procedures upon Steps 3 and 4

The most surprising feature of the change to the use of the percentile method for combining component boundaries in Step 3 was its lack of impact upon the observable behaviour of the awarders. In general, awarders did not appear to make allowance for the likely consequential changes to the aggregate standards when setting their component grade boundaries. Although these consequential changes were explicitly drawn to the awarders' attention, they nonetheless recommended component boundaries which led, as a result of the new percentile method, to more lenient aggregate boundaries. Indeed, there was a heated debate between the awarders and board officers in one of the Phase 3 meetings as to whether changes in component boundary standards were legitimate, even though all concerned agreed about the likely effect upon the aggregate standard of the new boundary aggregation method.

There is no way of knowing whether the awarders would have taken the same attitude if the change in boundary combination method had been such as to reduce the proportions of candidates awarded high grades, rather than increasing them. However, the discussion in Section 5.5 raises the possibility that the awarders' sanguine acceptance of the aggregate consequences of the new procedures might have been helped by the direction of their effect. However that may be, it certainly appeared that the relative complexity of the new procedure for combining component grade boundaries tended to discourage involvement by the awarders in Step 3. *You do the sums, and tell us what we've done* was not an untypical remark and suggests that the lack of transparency of the percentile method may be an issue. In general, the awarders chose to treat the change in boundary combination method (over which, of course, they had no control) as a technical matter of limited relevance to their task.

In terms of the analysis of boundary combination methods in Section 4.2.2, the implications of this are clear. The little account evidently taken by the awarders of the aggregate consequences of their component boundaries, implies that the use of a combination method which makes allowance for regression effects is required.

The impact on the behaviour of the awarders of the introduction into Step 4 of formal limits for annual changes in examination outcomes was much greater. This certainly appeared to increase, as intended, the extent to which account was taken of the statistical data (see Section 6.6 for further details). As in Step 2, the new arrangements made it easier for Subject Officers to ensure that the statistics were considered seriously at this stage. On the other hand, many awarders clearly saw the requirement for a written rationale simply as a manipulative device for making them change their recommendations to produce statistically acceptable outcomes. In two meetings, in particular, (Mathematics and Economics) some awarders explicitly argued against writing a rationale on principle and in favour, instead, of simply changing the boundaries to avoid having to do so. Although there is no necessary contradiction, it is interesting to note that one of the most committed critics of the new Step 4 procedure (*Whoever invented this* [the limits] *just doesn't understand how numbers work.*) was the individual (identified as A2 in the extract from the Phase 2 Mathematics award

discussed in Section 5.5.5) who argued very strongly for a decision based upon statistical evidence two years earlier.

Three years later, at the time of writing, further informal observations suggest that these more extreme reactions were essentially transitional responses to changes which were imposed upon the awarders by the board and which implicitly cast doubt on the validity of their previous well-established practice as groups of professionals. The more lasting observable effect of the new procedures in Steps 3 and 4 has been to reduce the extent to which awarders concern themselves with the final outcomes of the examinations. A greater concentration on Steps 1 and 2 is evident, after which complex calculations are done by the board's officers and the awarders are told of the results. Changes to boundaries in Step 4 are now much more common than under the procedures observed during Phases 1 and 2, but are largely done on the basis of the statistical evidence, rather than reconsideration of candidates' scripts. A similar relative lack of involvement of awarders with the final stages of their work was also reported by the recent (Murphy *et al*, 1996) study of GCSE awarding (which uses similar Step 3 and 4 procedures; see SCAA, 1995).

5.7 CONCLUDING REMARKS

The following conclusions from this chapter are particularly significant for the summative evaluation of awarding procedures given in Chapter 9:

1 In two important respects, awarders tend **not** to contextualise their qualitative judgements of candidates' work:

- they judge scripts largely as if they were independently constructed *artefacts*, rather than as *responses* to particular question papers;
- they judge work on each component with little regard for the cumulative demands of the examination as a whole.

2 Awarders have a strong commitment to being as fair as possible to the candidates, leading to a tendency to give the benefit of any doubt to the candidates by positioning grade boundaries at the lowest mark which they can justify.

3 When discussing grade boundary marks, awarders negotiate to achieve an agreement as close to their own view as possible, rather than discussing the evaluatively relevant characteristics of the work of candidates on the relevant mark points.

4 Awarding meetings exhibit many of the well-established features of small decision-making groups studied in other contexts: increased confidence and group polarization (the *risky shift*); pressure (including normative pressure) to agree; stereotypic perceptions of people, and devaluation of information, from outside the group; and so on.

5 Without clear procedural guidance to the contrary, awarders based their decisions almost exclusively upon their own professional judgements of individual candidates' work and were reluctant to modify or justify their decisions in the light of relevant statistical data.

6 Faced with a procedural imperative to give more weight to statistical evidence, some awarders tended to retreat into the process of forming qualitative judgements and took less interest in the overall outcomes of the examination. This tendency was exacerbated by the obscurity of the statistical process for determining aggregate boundaries from the awarders' component judgements.

7 Nonetheless, clearer specification of the tasks which must be carried out and their rationale, together with the encouragement of a distinct, facilitating, role by the Chair generally led to a more considered approach to the different sorts of evidence which are relevant to the positioning of the grade boundaries.

Evaluation of the outcomes of the awarding meetings and whether or not the procedural changes introduced between Phases 2 and 3 produced an improvement in practice, are the key issues addressed in detail in Chapter 6.

CHAPTER 6

A QUANTITATIVE ANALYSIS OF CONVENTIONAL AWARDING: THE OBSERVATIONAL WORK PART 2 AND OTHER DATA

"Judge not, that ye be not judged."
- *St Matthew, Chapter 7, Verse 1.*

6.1 INTRODUCTION

Following the qualitative analysis of the processes of awarding given in Chapter 5, this chapter takes a more quantitative approach. Two sources of data are used for this purpose: the encoded observations from Phases 2 and 3 of the observational work and the distributions of candidates' grades which summarise the outcomes of the awards.

6.2 THE PHASE 2 OBSERVATIONS OF AWARDING MEETINGS IN 1991

Phase 2 of the observational work primarily involved the systematic application of the coding scheme developed during Phase 1, with the aim of building a quantitative description of the nature of the judgements made by awarders. In the Phase 2 work, 7 award meetings were observed, all of which lasted for a single day. The subjects involved, in the order in which the meetings occurred, were:

1. General Studies,
2. Accounting,
3. Physics,
4. Economic and Social History,
5. Mathematics,
6. English Language and Literature,
7. Communication Studies.

In Phase 2, the observed meetings were chosen so as to represent the range of academic subjects examined at A-level. The meetings were audio tape recorded and their proceedings were encoded using the coding scheme developed in Phase 1. The tape recordings have already been drawn upon for the qualitative analysis of the awarding process presented in Chapter 5. In this section, the coded data from the Phase 2 meetings are analysed.

6.2.1 The coding scheme used for the phase 2 observations

As a result of the Phase 1 work, the coding scheme eventually used in the Phase 2 observations was considerably different from that originally devised and described in Section 5.4.4. It focussed on the nature of the awarders' discourse and the evidence they used to make their judgements. The categories were as follows:

+1	Positive affect/social
+2	Gives methodological guidance
+3	Gives evaluative criterion
+4	Gives procedural suggestion
+5	Gives overall judgement (candidates as a group)
+6	Gives evaluation of a particular script
+7	Gives statistical opinion or information
+8	Gives opinion or information concerning the paper
+9	Makes other relevant point
+0	Suggests boundary mark
-0	Asks for boundary suggestion
-9	Seeks other relevant information
-8	Seeks opinion or information concerning the paper
-7	Seeks statistical opinion or information
-6	Seeks evaluation of a particular script
-5	Seeks overall judgement
-4	Seeks procedural suggestion
-3	Seeks evaluative criterion
-2	Seeks methodological guidance
-1	Negative affect

As noted in Section 5.5.3, the plan for the Phase 2 observations included an analysis of the evaluative notes made by the awarders during Step 1. These data were used as the basis of an analysis of the nature of the evaluative judgements which the awarders made of the candidates' scripts. This analysis is reported in Section 6.3, below. There were, therefore, no distinct evaluative categories in the final coding scheme used for the observational work in Phase 2.

6.2.1.1 Applying the Phase 2 coding scheme

The use of a micro-computer to encode the discourse in the awarding meetings was tried and appeared to work well during Phase 1. The simple recording program, written in GWBASIC, which was developed during Phase 1 is given in Appendix 6.1. Using this system, it was possible separately to encode each individual participant's contribution to the discussion in the observed meetings. In this connection, it is worth noting that the 20 categories in the coding

scheme represent only 10 logical distinctions, each being coded either positively or negatively. The coding task was, therefore, easier than it might at first appear.

In any quantitative observational work, the reliability of the observations is obviously an issue. All the Phase 2 observations were carried out by the author so that inter-observer differences could not arise. As far as the internal consistency of the Phase 2 observations is concerned, subjectively this seemed to be high. Most of the categories in the coding scheme refer to speech acts with distinctive content which were easily identified. Moreover, although it has to be acknowledged that occasional lapses of attention could have resulted in some undercounting of these speech acts, the repeating structure of the meetings (see Chapter 5) gave a rhythm to the process of observation which helped considerably to maintain attention. Finally, indirect evidence for the reliability of the Phase 2 observations is available from the Phase 3 observations. These were carried out by two other observers, after training by the author, and quantitative data on the agreement between the Phase 3 observers, and between them and the author, are reported in detail later (Section 6.5.3.1). Here, it need simply be reported that no statistically significant inter-observer differences occurred in Phase 3.

As just noted, most of the categories in the Phase 2 coding scheme are reasonably self-explanatory. The way in which they were applied is described below.

> The *affective/social* category encodes remarks which were not directly concerned with the task of awarding but served to release tension, express annoyance, call people back to the task in hand and so on.
>
> The *methodological guidance* and *procedural suggestion* categories are fairly self-explanatory but may not be easy to distinguish. The former refers to remarks concerning the legitimacy of particular approaches or types of evidence; the latter encodes more immediate management of the meeting within the established procedures. For example, a proposal to look at more scripts would be coded as a procedural suggestion but a question as to whether it was sensible to ignore scripts which exhibited some evaluatively problematic feature would be coded as methodological.

The *evaluative criterion* category concerns explicit discussion of the criteria which should be used, or were being used, when evaluating scripts. As noted in Chapter 5, it was not possible or productive to encode, during the meetings, the occasional references to evaluative criteria which were made in passing when an individual script was evaluated.

The *overall judgement* category refers to judgements about the candidates as a group such as *I think the candidates are a little weaker this year.* The doubtful legitimacy of such judgements was discussed in Chapter 5.

The *evaluation of a script* category encodes evaluative judgements about individual scripts, whether or not reasons are given in support of the judgement. As noted in Chapter 5, supporting reasons are, in fact, only rarely given or discussed in most awarding meetings.

Statistical opinion or information covers both discussion of data about the percentages of candidates achieving particular marks and references to the absolute number (or proportion) of marks awarded to scripts.

The category for *opinion or information about the paper* was used to encode discussion of the qualities, such as difficulty, of the question paper as a whole or individual questions within it.

Finally, any proposal for a particular mark as the boundary being set was encoded as *a boundary suggestion.*

6.2.2 The focus of the awarders' discourse in Phase 2

The complete summarised raw data from the Phase 2 observations are given in Appendix 6.2. It can be seen from these data that the proportions of remarks involving the explicit seeking of particular contributions is relatively small and, for this reason the data for seeking or giving a particular sort of contribution have been combined in the analyses that follow.

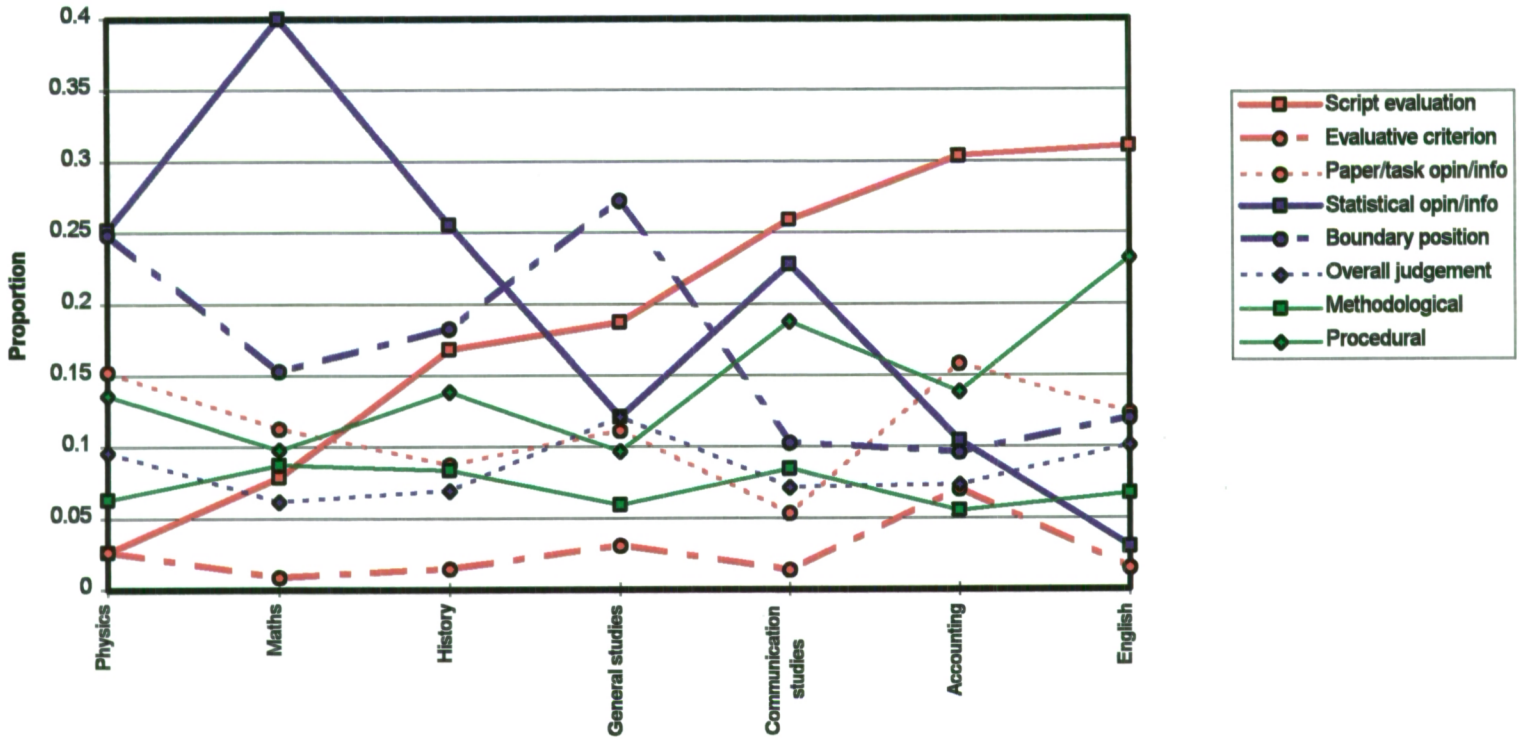
Figure 6.1 shows, for the meetings observed in Phase 2, the proportion of participants' remarks falling in each of the observational categories which is directly relevant to the task of awarding. (Affective/social contributions have been omitted from Figure 6.1; their frequency is reported in Appendix 6.2 and Figures 6.2 to 6.4.)

There are several features of Figure 6.1 worthy of particular note. First, the proportion of remarks relating to evaluative criteria is very low in all the meetings (the highest value being about 7% of remarks in Accounting). This is consistent with the qualitative analysis of Chapter 5 where the surprising lack of discussion about evaluative criteria was also noted.

Second, in Figure 6.1 the subjects have been ordered in terms of the proportion of remarks which gave an evaluation of a script (note that a single remark evaluating more than one script was encoded as several distinct script evaluations, one for each script referred to). It is apparent that there is a broadly inverse relationship between the proportion of script evaluations voiced in a meeting and the proportion of remarks referring to statistical data. Thus, at the extremes, in the Physics meeting about 25% of the discussion related to statistical data and less than 5% to individual scripts, whereas in English about 30% of the discussion was of individual scripts and less than 5% about statistics. (In the Mathematics meeting, over a third of the remarks made by participants related to statistical data but this arises, at least in part, from the particular difficulties of the Phase 2 Mathematics meeting and the attempt to find a solution which was described in detail in Chapter 5.) Third, the proportion of remarks involving suggesting a particular mark for the grade boundary also appears to be inversely related to the proportion giving evaluations of individual scripts (and positively related to the proportion about statistics).

Fourth, methodological and procedural discussion together account for about 20% of the remarks made in all the meetings, possibly implying a degree of uncertainty among the participants about their task. Methodological discussion is greatest in the meetings which emphasise the evaluation of scripts. As noted in Chapter 5, such discussion frequently centres upon the relevance, or otherwise, of statistical data to the task in hand. Fifth, around 10% of the awarders' discussion concerns overall judgements of the attainment of the

Figure 6.1
Proportion of observed participants' remarks in each category (excluding affective/social) - Phase 2



candidates as a group. The essentially speculative and irrelevant nature of such discussion was considered in Chapter 5. Sixth and lastly, about 10% of the remarks made in the meetings concern the question paper and could serve, therefore, to contextualise the decision making process.

The differences between the frequencies of different types of remark in the different meetings are highly statistically significant (χ^2 = 514, with 42 degrees of freedom). Table 6.1 shows the adjusted residuals (Everitt, 1977) from the expected frequencies. These can be interpreted as normal deviates, so cells with an absolute value greater than 1.96 are statistically significant at the 5% level.

Table 6.1
Adjusted residuals for frequencies of remarks in each category (excluding affective/social) - Phase 2.

Nature of remark	Physics	Maths	History	General studies	Communication studies	Accounting	English
Script evaluation	-7.82	-6.97	-1.13	-0.28	**2.60**	**7.90**	**5.13**
Evaluative criterion	-0.46	-3.09	-1.64	0.04	-1.58	**6.47**	-1.58
Paper/task opin/info	1.75	-0.64	-1.82	-0.64	-3.26	**3.20**	0.11
Statistical opin/info	**2.56**	**12.28**	**2.60**	-4.62	1.23	-6.58	-7.24
Boundary position	**4.10**	-0.81	0.78	**6.54**	-2.66	-5.31	-2.13
Overall judgement	0.80	-1.93	-0.93	**3.03**	-0.71	-1.09	1.06
Methodological	-0.49	1.65	0.98	-0.88	0.93	-1.58	-0.15
Procedural	-0.13	-2.80	0.03	-2.62	**2.24**	0.04	**4.70**

Note - **bold** and underlined text respectively indicate those positive and negative values which are statistically significant (< 0.05).

Thus, the Physics and Mathematics meetings discussed individual scripts significantly less than the other meetings and, along with the History meeting, discussed statistical data significantly more. At the other extreme, the Accounting and English meetings contained significantly less discussion of statistics and, with Communication Studies, significantly more discussion of the value of individual scripts.

A number of points need to be made about the interpretation of these data. First, the data relate to the frequency of remarks in the various categories. This is not, necessarily, the same

as the influence of those remarks in the decision-making process. In theory, therefore, it could be that a meeting discussed the value of many scripts but then made its decisions solely on the basis of a single consultation of the statistics. In such a case, the kind of data reported here would appear to imply a script-driven approach which was not actually adopted. However, it would seem perverse for groups of awarders, who were all experienced in their task, to spend large amounts of time in discussion which they knew to be irrelevant to their own decision-making process. Moreover, the picture which emerges from the data given here is broadly in line with the qualitatively rich account of the decision-making process given in Chapter 5 and, on this basis, it seems reasonable to argue that the differences between the meetings shown in Figure 6.1 and Table 6.1 genuinely reflect differing emphases in the types of evidence used by the meetings to reach their decisions.

The second interpretational point concerns the identification of the meetings with particular academic subjects. Is it the subject content which determines a meeting's emphasis on one source of evidence, in preference to another, or is it simply a matter of the particular approaches of different groups of people to the same task? It seems likely that the subject involved has some influence on the evidence emphasised by the awarders. As noted in Chapter 3, assessment in Mathematics and the sciences has, historically speaking, always been carried out largely by counting the proportion of questions which are correctly answered. In arts subjects, in particular English, the assessment process has always been much more evaluative. It seems likely, therefore, that the variations between meetings in the emphasis given to different types of evidence reflect, to some extent at least, the traditions of the subjects involved. Moreover, it is a reasonable hypothesis that question papers which consist of convergent answers to several problems make it more difficult for awarders to form an evaluative judgement of overall quality than papers which require candidates to develop an analysis in prose, at some length. Some evidence in support of this hypothesis is reported in Section 6.3 and the differences in approach between the Phase 2 meetings are consistent with it. Nonetheless, while it is sensible for the awarding approach to suit the academic subject concerned, it is hard to justify what happened in some of the meetings observed in Phase 2 where either statistical or qualitative information was almost completely ignored, rather than being considered and only then, perhaps, being discarded.

6.2.3 The different roles of the participants in the Phase 2 meetings

In this section, data on the differences between the contributions made to the meetings by the Chairs, board Officers and other participants (referred to hereafter simply as the *Awarders*) are reported. Table 6.2 reports the results of a χ^2 analysis of residuals on the frequencies of each category of contribution by different participants, when the data from all the Phase 2 meetings are combined. Overall, the differences between the frequencies for nature of remark by role were highly statistically significant ($\chi^2 = 670$ with 16 degrees of freedom).

Table 6.2
Adjusted residuals for frequencies of remarks in each category by role - all Phase 2 meetings.

Nature of remark	Chairs	Awarders	Officers
Script evaluation	1.23	**5.62**	-9.61
Evaluative criterion	-1.17	**3.72**	-3.69
Paper/task opin/info	-4.44	**8.21**	-5.65
Statistical opin/info	-3.35	-10.38	**19.23**
Boundary position	**7.52**	-2.52	-6.55
Overall judgement	-5.10	**6.60**	-2.49
Methodological/	0.39	-0.82	0.64
Procedural	**6.60**	-9.16	**4.10**
Affective/Social	-2.75	**2.51**	0.15

Note - **bold** and underlined text respectively indicate those positive and negative values which are statistically significant (< 0.05).

It can be seen that, on average, the Chairs and Officers make significantly more procedural contributions to the meetings than the Awarders. The Chairs also mention the positions of the boundaries significantly more often than the other participants, reflecting their job of bringing the meetings to decisions. The Awarders evidently concern themselves with the more qualitative aspects of the awarding process, referring significantly more often than the Chairs and Officers to evaluative criteria, the papers and tasks involved and the overall quality of the candidates. The Awarders also significantly more often evaluate the quality of individual scripts. The Officers make significantly more references to the statistical data. The data on

affective/social contributions are surprising because Chairs of meetings would not normally be expected to make significantly fewer such contributions than other participants. However, these data reflect the very atypical nature of one of the Phase 2 meetings, as will become clear shortly. Overall, the data in Table 6.2 are consistent with the expected roles of the participants in the meeting, although these were not well-defined by the Board's 1991 procedural documentation (see Appendix 5.1).

However, the aggregate data for all the meetings inevitably disguise differences between them. Table 6.3 shows the proportions of the contributions to each meeting made by Chairs, Awarders and Officers. The differences across the meetings in the frequencies of contributions by those with different roles are highly statistically significant (χ^2 = 198 with 12 degrees of freedom). Analysis of standardized residuals shows the Physics, History and English Chairs to have made significantly more contributions, and the Mathematics and General Studies Chairs significantly fewer, than the Communication Studies and Accounting Chairs. The Officers in the Mathematics and History meetings made significantly more, and the Officers in Accounting and English significantly fewer, contributions than their peers in the other meetings.

Table 6.3
Proportions of contributions to each Phase 2 meeting by chairs, awarders and officers

	Physics	Maths	History	General Studies	Commun. Studies	Account.	English
Chairs	0.43	0.30	0.48	0.24	0.37	0.32	0.42
Awarders	0.40	0.49	0.29	0.60	0.48	0.61	0.52
Officers	0.17	0.21	0.23	0.16	0.15	0.07	0.05

Figures 6.2 to 6.4 show, separately for the Chairs, Awarders and Officers respectively, the proportions of remarks in each category. From these graphs, several points emerge. Although, compared with the Awarders, a greater proportion of the Chairs' contributions are procedural, as is to be expected given their role, there is otherwise a marked qualitative similarity between the emphases on qualitative and statistical data in the remarks made by each Chair and the emphases for each meeting as a whole (compare Figures 6.1 and 6.2).

Figure 6.2

Proportions of remarks in each category from chairs in Phase 2 meetings

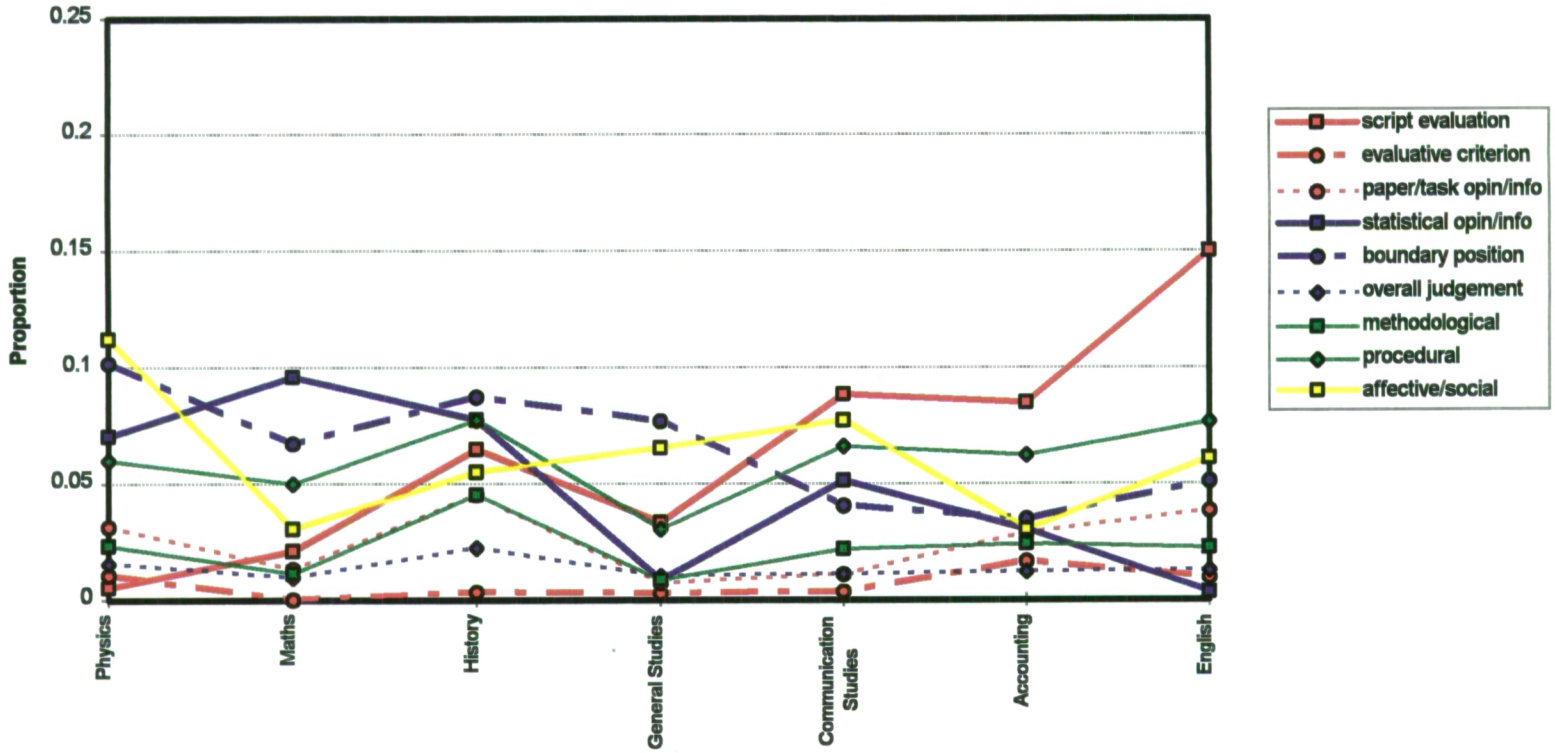


Figure 6.3

Proportions of remarks in each category from awarders in Phase 2 meetings

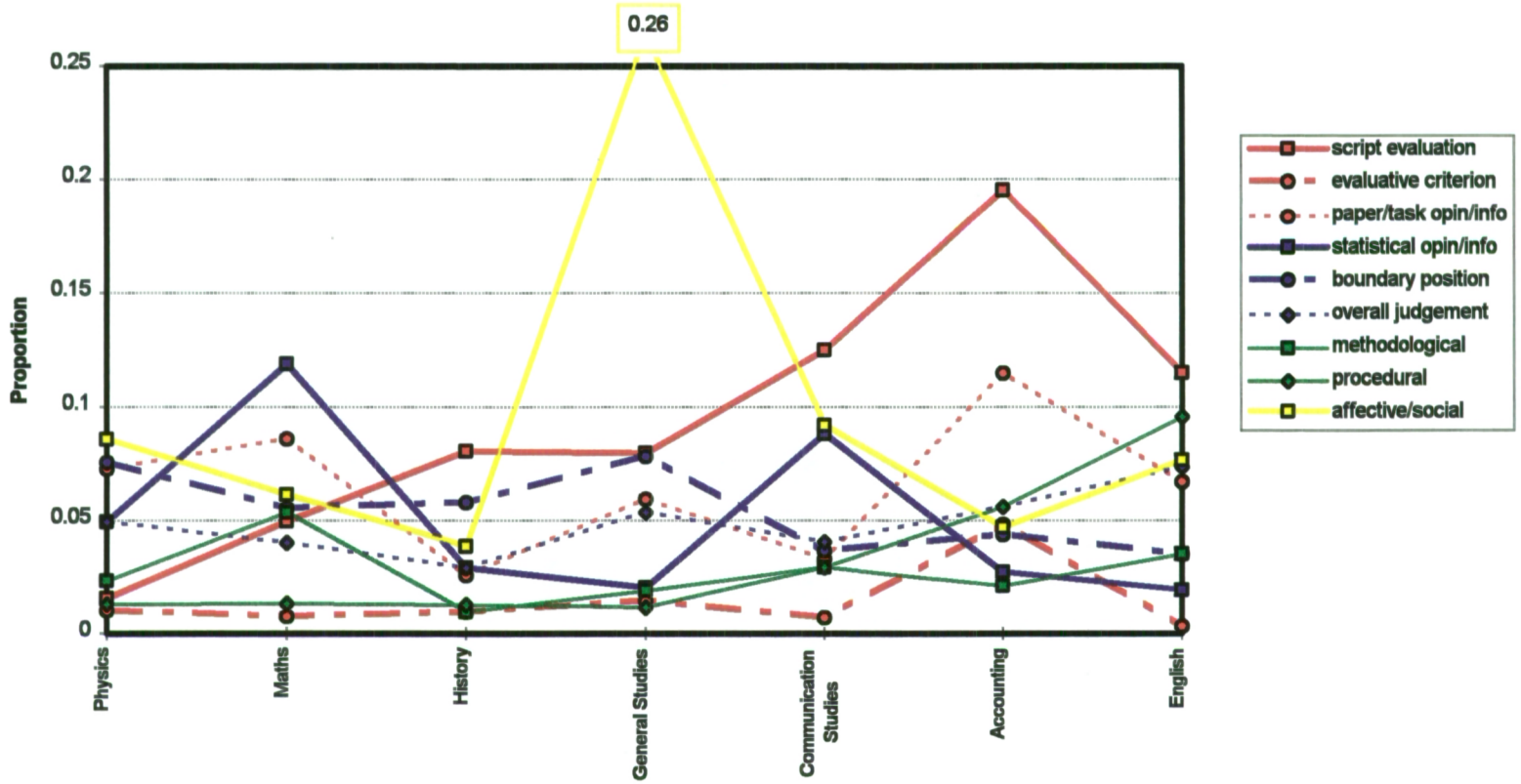
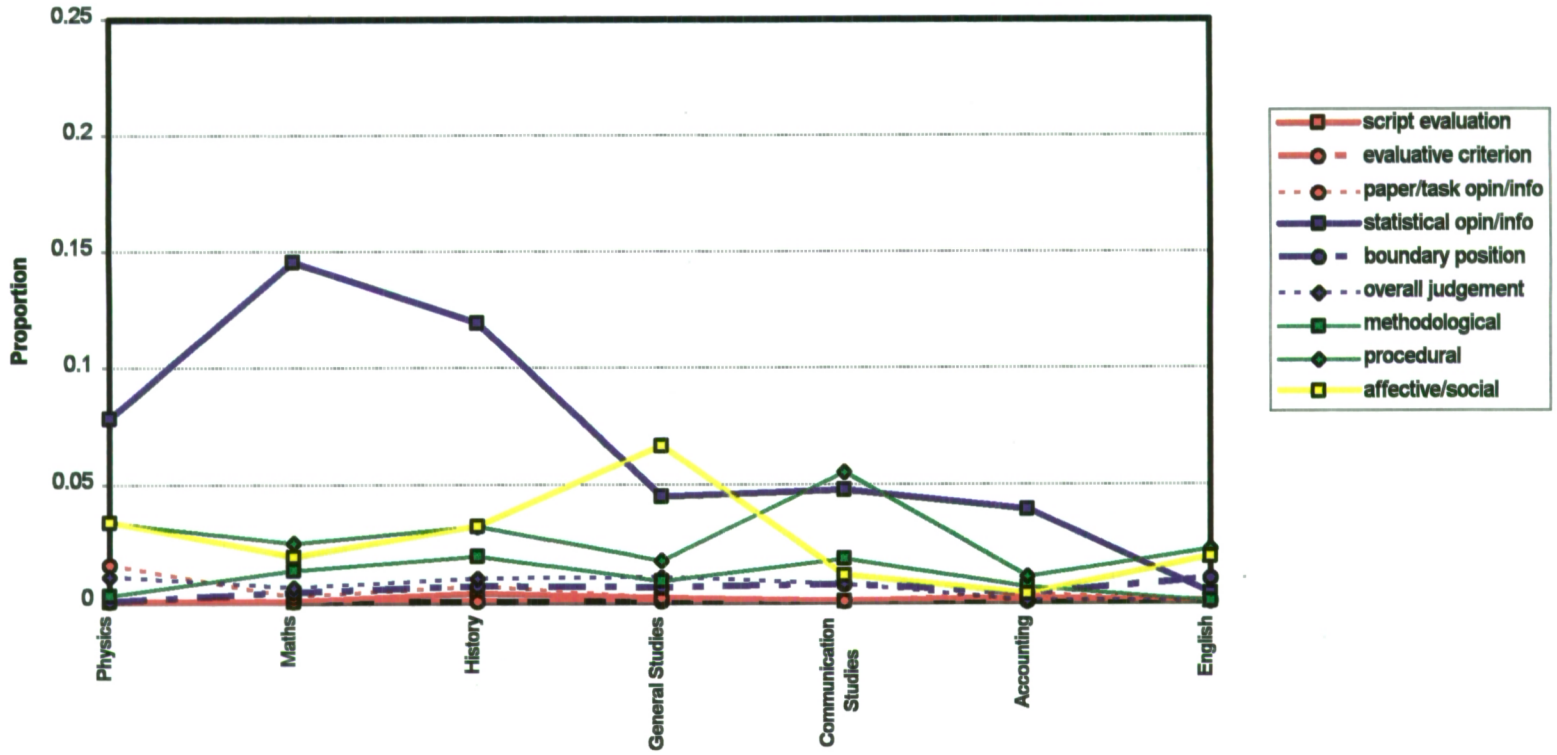


Figure 6.4

Proportions of remarks in each category from officers in Phase 2 meetings



It is not possible to be certain from the Phase 2 work alone whether the similarity between the emphases on qualitative and statistical data in the remarks made by each Chair and the emphases for each meeting as a whole reflects the influence of the Chairs upon the tone of the meetings or their response, in common with the other awarders, to the nature of assessment in the subjects concerned. This point will be considered again in Section 6.5.5 when the corresponding analysis of the Phase 3 observations is discussed. The very high proportion of affective/social remarks from the Awarders in the General Studies meeting is noteworthy and consistent with the significantly low proportion of contributions made by the Chair to this meeting. Finally, the Officers' contributions to the meetings also vary across subjects and, although generally of a statistical nature, clearly reflect the prevailing preference of each meeting for statistical or qualitative evidence.

It thus appears that the preference of each meeting for particular sorts of evidence is reflected in the behaviour of all the participants. In the case of the Officers, this is surprising because, as reported in Chapter 5, they were generally seen by the Awarders as having different concerns. It appears, however, that either the Officers shared the views of the Awarders on the relevance of statistical or qualitative data, or they were affected by normative pressure to conform to the approach of the meeting. Strength is given to this latter hypothesis by the observation that the two meetings where the Officers made significantly few contributions (Accounting and English) were those where the Chairs and Awarders placed most emphasis on script evaluation. The qualitative similarity between the contributions of the Chairs and Awarders implies that the role played by the Chair was not clearly delineated in most of the meetings. Finally, both this and the preceding section, reveal considerable variation in the approaches taken to the awarding task by the different meetings.

6.3 AWARDERS' REASONS FOR THEIR EVALUATIVE JUDGEMENTS

From the Phase 1 observations reported in Chapter 5 it rapidly became clear that, during the discussion at the meetings, awarders rarely gave explicit reasons in support of their evaluations of individual candidates' scripts. As a result, it was decided not to attempt to gather data on the reasons awarders have for their evaluations during the observational work

in Phases 2 and 3. However, during the Phase 2 meetings the awarders were asked to make notes on each script which they evaluated and these provide a source of data on the reasons which they have for their evaluative judgements. These data are reported in this section.

6.3.1 Coding the awarders' evaluative reasons

There was a full discussion of the classification of reasons for evaluative judgements in Chapter 5 (Section 5.4) but the taxonomy used for the analysis reported in this chapter was, as with the observational categories, modified in the light of the observations. Those modifications and the final coding scheme for evaluative reasons are described in this section.

6.3.1.1 Affective reasons

In analysing the awarders' notes on individual scripts, a new, and frequently occurring, sub-class of affective reasons was identified. This is exemplified by simple non-specific responses to scripts such as "good", "borderline", "weak" and so on. It is certainly doubtful whether these should be called **reasons** but they are characteristic of the justifications which some examiners give for their evaluations of candidates' work (for example, *Just not quite good enough to be an A*). As with the more specific affective reasons discussed in Section 5.4.2, these are not formally **reasons** for the evaluation, but frequently have to be accepted as the best **explanations** available.

6.3.1.2 Other types of reasons

In addition to the types of evaluative reasons discussed in Section 5.4, which were based upon theoretical considerations in a related field, the relevance of two other categories was apparent from awarders' notes. The first concerns candidates' marks and is necessary because awarders sometimes refer explicitly to the number of marks awarded to a script in order to justify their judgement. For example: *for a pass, there's too little accomplishment - 33% - in Section A.* The second category relates to the extent to which a candidate's performance is balanced (or consistent) throughout the script. For example: *only three questions were answered well, the rest were poor.*

6.3.1.3 The operational coding scheme

Based upon the analysis in Chapter 5, modified as just described, the following categories were used to encode the types of reasons given by the Phase 2 awarders for their evaluative judgements of candidates' work:

Genetic reasons
Moral and Social reasons
Affective reasons
Stylistic Unity, Structure and Complexity
Script Content
Marks
Script Balance

The awarders' reasons were cross classified, under the above types, as:

holistic or fragmented; and
contextualised or uncontextualised.

Despite the difficulties of interpretation affecting these last two dimensions, in practice it was not difficult to categorise the recorded reasons as holistic or fragmented. Moreover, in their recorded reasons, the awarders always referred only to the work in the script, making no reference to the difficulty of the questions which had stimulated it. All their reasons, as recorded, therefore appeared uncontextualised. The interpretation and implications of this are discussed later.

A copy of the form upon which the awarders at the Phase 2 meetings gave their reasons is attached as Appendix 6.3. Since the awarders were free to write anything they felt appropriate, they sometimes gave more than one reason for a single script. Where this happened, their responses were counted in every relevant category. Thus, for example, comments such as *good script with no problem areas* were coded twice: once in the *Affective* category (*good script*); and once in the *Script Balance* category (*no problem areas*). Many of the non-specific affective responses were, as in this example, given in combination with a more specific reason and this should be borne in mind when interpreting the data reported in the next section.

6.3.2 The nature of the reasons

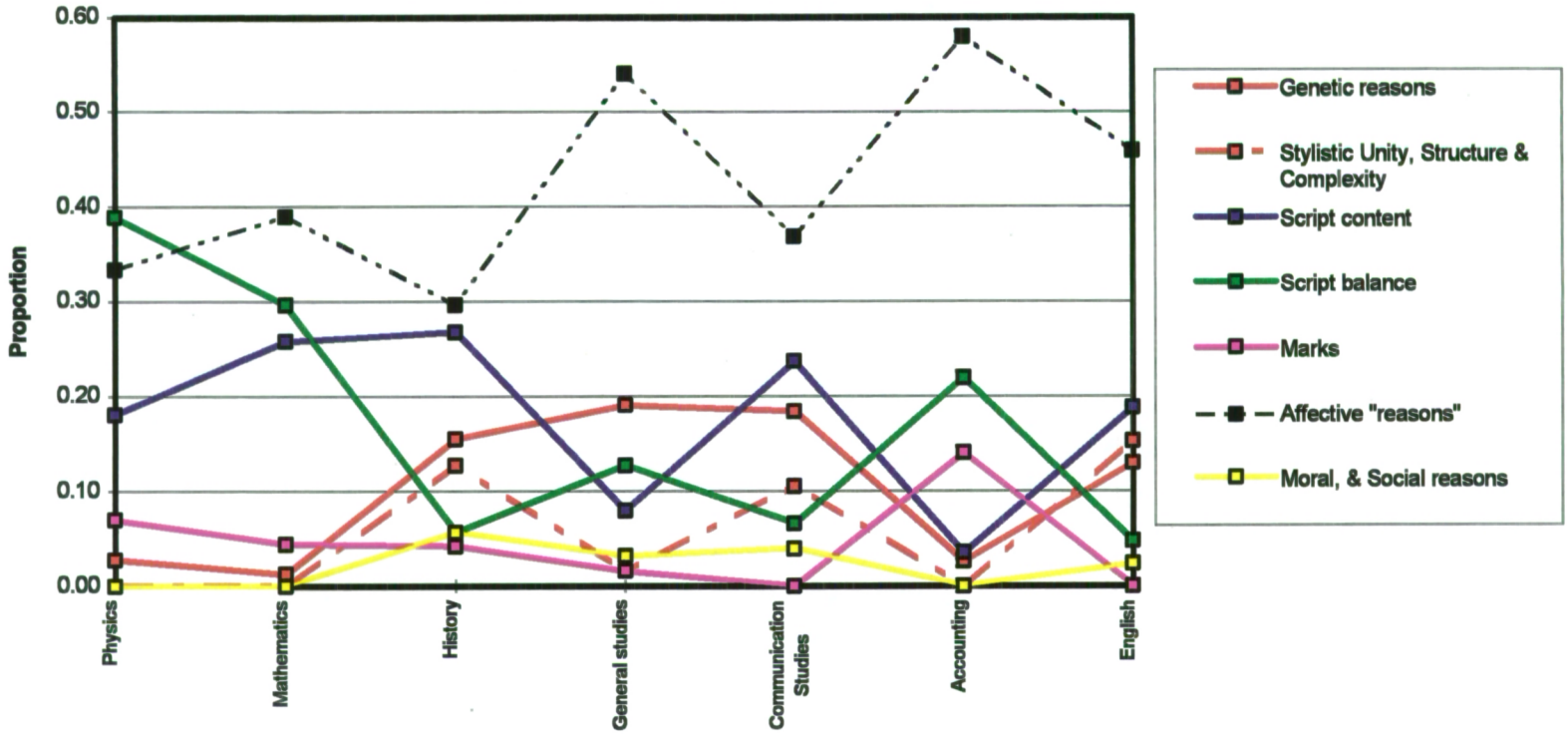
Appendix 6.4 contains the raw data showing the frequency of each type of reason for an evaluative judgement given by the Phase 2 awarders. Across all the meetings, the proportion of holistic reasons is 0.83. Since the job of the awarders is to judge each script which they evaluate as a whole, the remaining 17% of fragmented reasons must be considered inadequate (see Chapter 5). However, the difficulties of interpreting the holistic/fragmented distinction were discussed in Section 5.4.2.7. In essence, it is impossible to know when an awarder offering a fragmented reason is simply leaving as implicit their satisfaction or disatisfaction with the remainder of the script. It seems safe to conclude that particular fragmented features of scripts sometimes override more holistic considerations when awarders are making evaluative judgements but, on the present evidence, it is possible only to say that this occurs, on average, in no more than 17% of such judgements. In the remainder of this section, the distinction between fragmented and holistic reasons has been ignored and only the evaluative type of the reason is considered, using combined data.

Figure 6.5 shows the proportions of the reasons, offered by the awarders at each Phase 2 meeting for their judgements, which fall into each type. From this figure, it can be seen that there are quite large differences between the subjects in the nature of the reasons given in support of evaluative judgements. As with the types of remarks made in the meetings, the question arises of whether these differences reflect the nature of the academic subjects concerned or are simply the established traditions of different social groups. The answer, too, is the same. It will be argued below that the differences observed are consistent with the traditions of assessment in the subjects concerned and the nature of the examination papers used in the different subjects.

A second question is whether the reasons given by the awarders to support their judgements reflect the criteria they actually use to make those judgements. It seems reasonable to argue that they reflect the type of criteria which the awarders **believe** they use because the notes from which they are taken were used by the awarders to remind themselves of their views about particular scripts. Whether the awarders are mistaken about their own criteria, or overestimate the importance of some criteria at the expense of others, is unknown. The well-

Figure 6.5

Proportion in each category of observed awarders' reasons for their evaluative judgements - Phase 2



known work of Nisbett and Wilson (1977) shows that such possibilities are real although White (1988) later offered a more optimistic view. Nonetheless, the present data can, at least, be interpreted as indicating the types of features of candidates' works which awarders believe to be important in arriving at the evaluative judgements involved in grade awarding.

The data illustrated in Figure 6.5 have been analysed using the same approach as in Section 6.2.2. The differences between the frequencies of different types of reason in the different meetings are highly statistically significant (χ^2 = 224, with 36 degrees of freedom). Table 6.4 shows the adjusted residuals (Everitt, 1977) from the expected frequencies. These can be interpreted as normal deviates, so cells with an absolute value greater than 1.96 are significant at the 5% level.

Table 6.4
Adjusted residuals for frequencies of reasons in each category - Phase 2.

Type of reason	Physics	Maths	History	General studies	Communication studies	Accounting	English
Genetic, reasons	-1.87	-3.81	**2.20**	**3.12**	**3.26**	-2.51	1.54
Stylistic Unity, Structure & Complexity	-2.03	-3.28	**3.26**	-1.27	**2.46**	-2.66	**4.82**
Script content	-0.02	**2.89**	**2.00**	-2.21	1.34	-4.47	0.18
Script balance	**4.60**	**3.96**	-3.03	-1.33	-2.92	0.91	-3.59
Marks	0.80	-0.40	-0.32	-1.31	-2.13	**4.88**	-2.27
Affective "reasons"	-1.73	-1.12	-2.39	1.88	-1.12	**3.59**	0.61
Moral and Social reasons	-1.19	-1.92	**2.69**	0.94	1.59	-1.56	0.48

Note - **bold** and underlined text respectively indicate positive and negative values which are statistically significant (< 0.05).

Two points about the reasons given by the Accounting awarders, in particular, are worth making. First, Figure 6.5 casts an interesting light on the superficially surprising similarity between the data on the nature of the awarders' discourse in Accounting and English (Figure 6.1 and Table 6.1). The present data show that, although both of these meetings involved a large proportion of explicit evaluations of individual candidates' work, those evaluations were justified on quite different bases. In particular, the Accounting evaluations were justified

largely by reference to the balance of candidates' scripts, measured in terms of the marks they earned. Second, while the low proportion of reasons mentioning content in General Studies seems consistent with the nature of the subject, the low proportion in Accounting is surprising. It may be that the Accounting awarders replaced explicit mention of the qualities of the scripts in terms of their subject content with reference to the marks which they had gained. To the extent that the marking scheme in Accounting involved awarding marks for points correctly covered, this seems a reasonable substitution.

One other point of detail seems particularly interesting. The History awarders gave significantly more Moral and Social reasons than the others. History was one of the two subjects identified in Section 5.4.2.2 where Moral and Social reasons seem most legitimate.

Excepting General Studies and Accounting, the awarders gave about 20% of their reasons for evaluative judgements in terms of subject content and about 40% in terms of affective response. This leaves about 40% of the reasons either to considerations of balance of performance within the scripts, to considerations related to the genetic, moral and social characteristics of scripts, or to the stylistic unity, structure and complexity of the candidates' work.

As with the data on the focus of awarders' discussions, it is possible to identify two extreme types of awarder. One extreme gives reasons related to the balance of candidates' performances throughout the script, (Physics, Mathematics and Accounting) measuring this sometimes in terms of the proportion of relevant subject content known (Physics and Mathematics) and sometimes in terms of the marks awarded (Accounting). The other extreme gives primarily genetic reasons, supported (for History, Communication studies and English) by reasons relating to the unity, structure and complexity of candidates' work. Broadly speaking, these two styles of justification correspond, respectively, to examinations involving the solution of several distinct problems and to examinations involving essay writing.

An important feature of the data reported in this section is the high proportion of affective reasons (more properly, *explanations*) given by the awarders in every subject. Moreover,

these affective responses appear particularly common in the absence of specific reasons relating to subject content. As noted in Chapters 3 and 5, the frequent use of affective justification is consistent with the idea that awarders' evaluative judgements have much in common with judgements of aesthetic value.

Finally in this section, the implications of the complete absence of contextualisation of the awarders' reasons need discussion. It could be, of course, that the judgemental process was a contextualised one but that the awarders felt this to be so implicit as not to be worth noting. However, given the relatively small amount of discussion (10% of observed remarks) of the question papers in the observed meetings (see Section 6.2), it seems reasonable to take the data on contextualisation at face value to some extent, at least. Such an interpretation is also consistent with the nature of the Chief Examiners' reports observed in Step 0 of the awarding process during Phases 1 and 2. These reports focussed mainly upon the attainment of the candidates, rather than evaluatively relevant features of the current year's question papers (see Section 5.5.1). The conclusion, therefore, is that the awarders made their evaluative judgements of the quality of candidates' work with little reference to the difficulty of the tasks which the candidates were trying to do. Unfortunately, this approach, the Script as Artefact rather than Script as Response awarding strategy, is inappropriate for the task of maintaining standards from year to year (see Chapter 3).

6.4 THE OUTCOMES OF THE 1991 MEETINGS

When awarders use the Script as Artefact awarding strategy, the observed changes in the outcomes of their meetings from one year to the next will not only include any changes in the attainment of the candidates between years, but also the effect of differences in difficulty between the two years' examinations. This is, of course, why the Script as Artefact strategy is inappropriate. Thus, considering the changes in outcomes between 1990 and 1991 - the years of the Phase 1 and Phase 2 observations - casts light on the adequacy of the awarding process. Can the scale and nature of the changes in outcomes reasonably be explained only by changes in the attainment of the candidates or is there reason to believe that effects due to the examination and its awarding are also present? This section addresses this question.

However, it is important to be clear about the nature of the investigations reported here. In the absence of independent information about the attainments of the candidates, it is impossible to disentangle the effects of candidate attainment and examination difficulty within any observed changes in a particular examination's statistics. This section therefore considers the balance of probabilities, based upon an investigation of many different examinations. This essentially statistical approach enables the existence to be demonstrated, with a high level of probability, of changes in examination outcomes which are due to the examinations, rather than the candidates. It does not, however, enable the particular examinations in which such effects operate to be identified.

Table 6.5 shows the changes which occurred between Summer 1990 and Summer 1991 in the cumulative percentages of candidates awarded Grades A, B and E in those of the board's Mode 1 A-Level examinations which attracted over 500 candidates in 1991. Some examinations exhibit small changes in overall outcomes between the two years, some examinations exhibit large ones. One point immediately worth making is that there is no obvious relationship between the subjects and the scale of the changes. Specifically, subjects which were shown to emphasise different forms of evidence by the Phase 2 observational work do not, as a result, differ consistently in the stability of their outcomes.

In Step 4 of the awarding process (see Section 5.3.5), the awarders are asked to explain any large changes in the proportions of candidates awarded each grade. In these circumstances, three possible explanations, one or more of which was usually offered by the awarders, were observed (see Section 5.5.5). In *Explanation Number 1*, the awarders simply argue that the entire group of candidates, as such, is better (or worse) than the previous year's group. In *Explanation Number 2* they go a little deeper and refer to any changes there may have been in the relative proportions of candidates entered by different types of centres or in the gender balance within the entry. (Descriptive information about the composition of the entry is routinely available to awarding meetings in terms of these two variables.) *Explanation Number 3* concerns the case where the number of candidates entering for the examination has changed considerably. It may then be suggested by the awarders that the candidates who have been gained or lost, as a group, are better (or worse) than the rest of the candidates.

Table 6.5
Comparisons between the outcomes in 1990 and 1991 for A-level examinations with more than 500 candidates in 1991

Subject	Number of cands in 1990	Number of cands in 1991	Change in cum % at Grade A	Change in cum % at Grade B	Change in cum % at Grade E
Accounting	6775	6637	-1.4	-3.9	-0.1
Applied Mathematics	1359	1237	-6.5	-3.4	0.1
Biology I	4152	5159	-4.2	-9.4	-12.4
Biology II	2195	2464	-5	-8.4	-12.7
Business Studies	11206	12477	1.3	3.2	6.2
Chemistry	3902	4139	-2.7	-5.8	-8
Communication Studies	3945	4565	2.5	6.4	2.1
Computing	3445	3294	-2.6	-4.2	-0.7
Constitutional Law	701	669	0	0.8	13.6
Economic & Social History	928	1066	-1.2	-2.6	-4.4
Economics	12056	11913	0.1	-0.8	-0.9
English I (Language & Literature)	11196	12186	-0.3	-2.1	-0.9
English II (Literature)	4401	3680	-4.8	-11.4	0.1
English III (Literature Alternative)	10061	13929	4.2	0.9	-3.6
Environmental Science	562	794	-0.2	0.4	1.4
French	4492	5324	1.2	5.4	5.1
General Studies	1141	1256	2.8	6.6	8.9
Geography	2813	3026	1.4	8.8	9
German	1706	2140	-1.6	-7	-10
Government & Politics	1658	1835	0.1	-2.9	1.8
History	5234	5894	0	-2.2	-4.2
History (Alternative)	2125	2397	-0.6	-1.6	-3.4
History of Art	1233	1226	1.9	8.5	6
Human Biology	2832	3300	0	0.1	0.4
Law	3747	4166	-4.2	-3.8	-0.1
Philosophy	476	699	-3.8	-4.1	-15.3
Photography	1335	1307	0.9	0.4	0
Physics	7412	7423	0	-2.8	-5.2
Physical Education	222	630	0.4	1.6	-1.1
Psychology	8504	10120	1.1	1.2	4
Pure & Applied Mathematics	6718	6647	-0.3	0.6	-0.2
Pure Mathematics	3219	3137	-13.2	-15.7	-4.4
Pure Mathematics & Statistics	5464	5432	0.4	1.6	-1.4
Sociology	19789	17222	0.5	-2.7	0.7
Spanish	773	928	-4	-11.8	-10.7
Sport Studies	477	749	-0.3	-1.8	2.4
Statistics	1939	1778	2	9.8	9.9
Theatre Studies	4749	5634	-0.9	-0.5	-1.4

In the following sections, each of the awarders' three explanations is examined, in the context of the data in Table 6.5.

6.4.1 Explanation Number 1: As a whole, the candidates as a group are simply better (or worse) this year.

As noted in Chapter 5, by itself, this is not an explanation at all for a change in the proportions of candidates in each grade. It is simply a restatement of the implications of the grade boundary decisions in a different form. Since it is offered by the same individuals who have made the grade boundary decisions which it purports to explain, it is not independent corroboration of the results of those decisions and cannot, logically, explain them.

However, this explanation is interesting because of the implied models which are held by those who proffer it. The explanation could be based upon the notion that systematic overall attainment changes are to be expected because of changing educational policy and practice or other external factors; or it could be the reflection of an implicit assumption that some variation is to be expected between the results of adjacent years' candidates simply because they are different groups of students; or it could be referring to both of these potential causes for variations in examination outcomes. The second cause will be considered first. Given that it is clearly a possibility, the obvious question to ask is: how large are the random variations in the statistics of public examination results which can be expected? This question effectively views each year's candidates for a particular examination as a sample from the population of all candidates who take that examination over its lifetime. Posed like this, it is essentially a question for sampling theory.

Sampling theory generally assumes that samples are randomly drawn from a well-defined population. However, defining the population from which successive years' candidates for a particular examination are drawn is not easy. Suppose, for example, that a different examination in the same subject ceases to be offered and some of the candidates who would have taken it now take an alternative examination. This can be viewed as adding more candidates to the population relevant to the alternative examination **after the sample for the previous year has been drawn**. To make progress, therefore, it is necessary to assume that changes of this type in the "population" do not occur. For the purposes of this section,

however, this assumption in not unreasonable because the existence of "population" changes provide the rationale for Explanation Numbers 2 and 3 which are discussed later. Explanation Number 1 does not refer to such changes so it seems legitimate, when evaluating it, to assume that the group of candidates who take an examination throughout its lifetime **can** be treated as a well-defined population from which each year's entry is a sample.

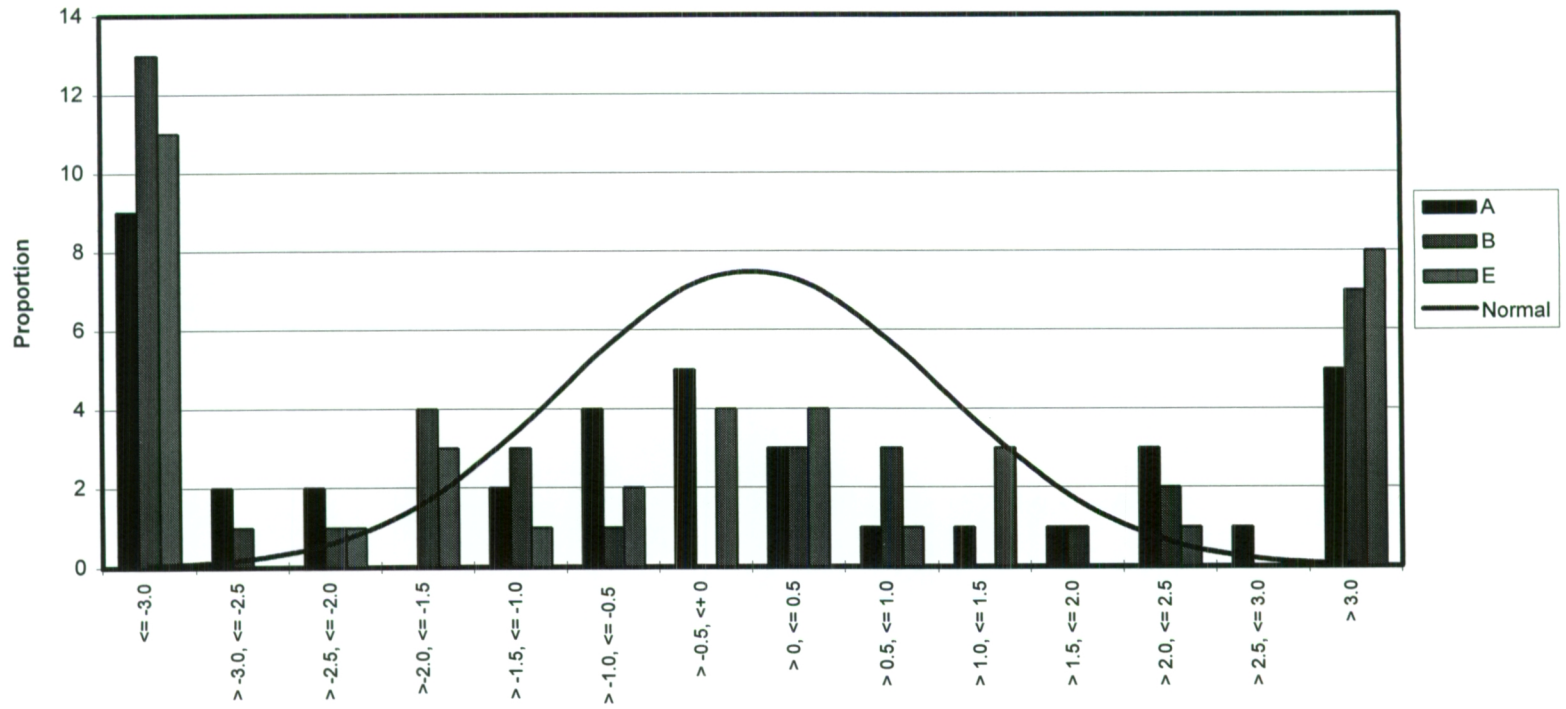
The other major assumption needed before proceeding further is that each year's candidates can be treated as a **random** sample from the notional population defined above. Given the age-related nature of entry to public examination, each year's entry is clearly not strictly random. However, for the purpose of evaluating Explanation Number 1, the high degree of chance involved in human conception means that it does not seem unreasonable, in the absence of any known external factors (such as an epidemic affecting pregnant mothers in the year in question), to treat any one year's age cohort as a random sample of independently chosen individuals.

If these arguments are accepted, then it is straightforward to evaluate the size of the differences in examination results statistics which might be expected as a result of chance differences between successive years' entries. The standard test for the significance of differences between proportions in large samples can be used (see, for example, Guilford and Fruchter, 1973). This test provides a statistic, z, which is theoretically normally distributed with a mean of 0 and variance of 1. If it is applied to the data producing Table 6.5 (these data are given in detail in Appendix 6.5), the results summarised in Figure 6.6 are obtained. Clearly, the differences in outcomes between 1990 and 1991 cannot reasonably be viewed as the results of random variations between successive groups of candidates.

It follows that some effect other than straightforward sampling error is operating in many of the cases in Table 6.5. It might be argued that the prior ability or motivation of the groups of candidates entered for the different examinations changed significantly between 1990 and 1991 for some systematic reasons. Possible causes of such a change might include, for example, widespread medical factors or social ones such as growing fear of unemployment, although such extrinsic factors seem unlikely to affect different school subjects differentially,

Figure 6.6

Distribution of z statistics for differences between outcomes in 1990 and 1991 for each key grade



as they would have to do to explain the data in Figure 6.6. Alternatively, in the terms of Explanation Number 1, the causes of the changes in outcomes between 1990 and 1991 must be improvement or deterioration in the quality of educational provision in the subjects concerned. Thus, to evaluate Explanation Number 1, the key question is whether changes in the overall ability or motivation of candidates or in educational provision could be the causes of the significant changes in examination outcome illustrated in Figure 6.6.

With the data available, this question cannot be unequivocally answered. However, evidence relevant to it can be obtained by considering the changes in outcomes for subgroups of candidates whose origins are different and whose educational provision is differently organised. This has been done for the examinations in Table 6.5 by looking separately at the changes in outcomes between 1990 and 1991 for UK candidates from schools, UK candidates from further education colleges and overseas candidates (the relevant data are contained in Appendix 6.5). In all, there are 67 cases in Table 6.5 where there is a significant difference between 1990 and 1991 in the cumulative proportion of all candidates at a key grade boundary. In 47 of these cases, a change in the same direction occurs for all three subgroups of candidates and, in the remaining 20 cases, such a change occurs for 2 out of the three subgroups.

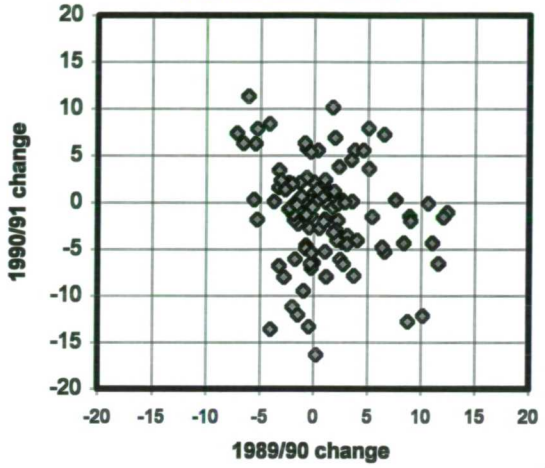
Moreover, z tests have also been done for the changes in outcomes in the same subjects between 1989 and 1990 (the relevant data are given in Appendix 6.5) and, in Figure 6.7, the results are compared with those illustrated in Figure 6.6. The annual changes in outcome between these adjacent pairs of years are clearly little related to each other.

It is difficult to identify any plausible extrinsic factors or educational mechanisms which could not only differentially affect overall attainment in different subjects on a global scale so markedly, but also produce effects which vary so much from one year's cohort of candidates to another. On balance, therefore, Explanation Number 1 appears insufficient to explain the scale of changes observed in the proportions of candidates awarded each grade between 1990 and 1991. Any mechanism capable of generating the changes observed would have to operate through something which all subgroups of candidates within an annual cohort have in

common but which changes annually. The examination itself is the only known factor which meets these requirements.

Figure 6.7

Relationship between z statistics of changes in outcomes for 1989/90 and 1990/91



6.4.2 Explanation Number 2: The balance of centre types and/or genders has changed.

This explanation seems at first sight to be a plausible one. Clearly, if there are differences in attainment between different subgroups of candidates, then variations in the relative proportions of these subgroups will lead to changes in the overall proportions of candidates awarded each grade. Is this effect sufficient to explain the differences observed in Table 6.5?

To explore this question, the grade distributions for subgroups of candidates in 1990 (see Appendix 6.5) were combined, re-weighted in such a way as to reflect the relative proportions of each subgroup in 1991, as follows:

$$P'_x = \sum_j P_{jx} \cdot s'_j$$

where P'_x is the predicted proportion of 1991 candidates exceeding the boundary for Grade x,
P_{jx} is the proportion of 1990 candidates in Subgroup j exceeding the boundary for Grade x
and s'_j is the proportion of candidates in Subgroup j in 1991

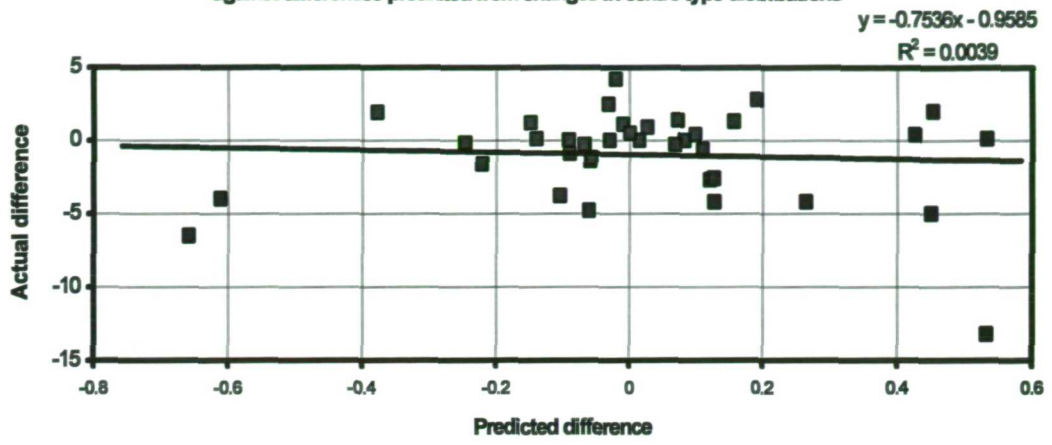
The changes in outcome predicted in this way were then compared with the actual changes in overall grade distribution between 1990 and 1991. Clearly, any differences between the actual changes in outcomes and the ones predicted by re-weighting indicate discrepancies which cannot be accounted for by changes in the composition of the entry between the two years, at least with respect to the subgroups referred to in Explanation Number 2. The results of these analyses are shown in Figures 6.8 and 6.9 and it is apparent that Explanation 2 does not explain the differences in outcome between 1990 and 1991. The observed differences are not only very much larger than those predicted but also uncorrelated with them. Appendix 6.6 shows why, in general, realistic changes in subgroup distributions are unlikely to produce large changes in overall examination oucomes.

It is sensible to ask at this point if there might not be differently attaining subgroups of candidates, other than those identified by the Board's operational data, which might occur with a frequency which varies substantially between years and therefore cause some of the observed changes in the proportions of candidates awarded each grade. It is not possible to rule this possibility out with complete certainty but it seems unlikely that any such effect is large. There is no reason to believe that examination candidates have among them subgroups which, independently of gender and centre type, vary substantially with respect to prior attainment **and in their incidence from year to year**. It could even be argued that monitoring the composition of the entry for an examination in terms of school type provides a crude surrogate measure for another strong correlate of examination success: socio-economic status. In fact, year-on-year variations in the proportions of candidates coming from schools of different types are small, being rarely greater than 3% of the total entry in large entry subjects.

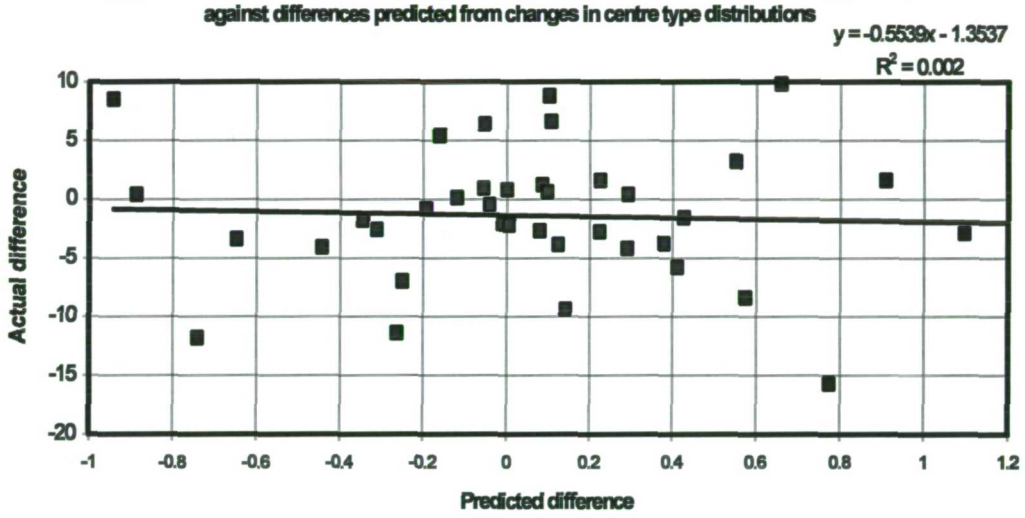
In any case, a very large proportion of the candidates for any particular examination in one year is normally entered by centres who also entered candidates in the preceding year. For the examinations in Table 6.5, Figure 6.10 shows the distribution of the proportions of candidates entering a particular examination in 1991 who came from centres entering candidates for the same examination in 1990. The mean proportion is 88%.

Figure 6.8

Actual differences between cumulative percentages of candidates at Grade A in 1990 and 1991, against differences predicted from changes in centre type distributions



Actual differences between cumulative percentages of candidates at Grade B in 1990 and 1991, against differences predicted from changes in centre type distributions



Actual differences between cumulative percentages of candidates at Grade E in 1990 and 1991, against differences predicted from changes in centre type distributions

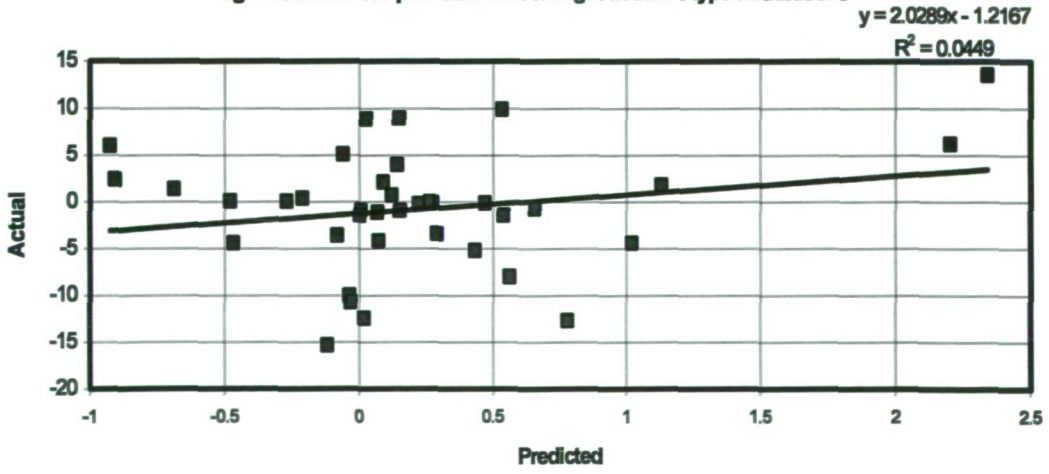
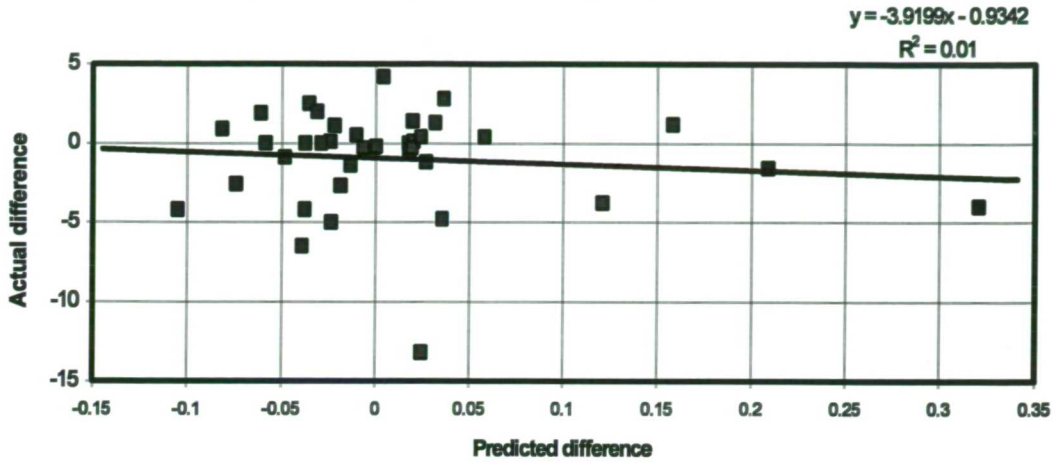
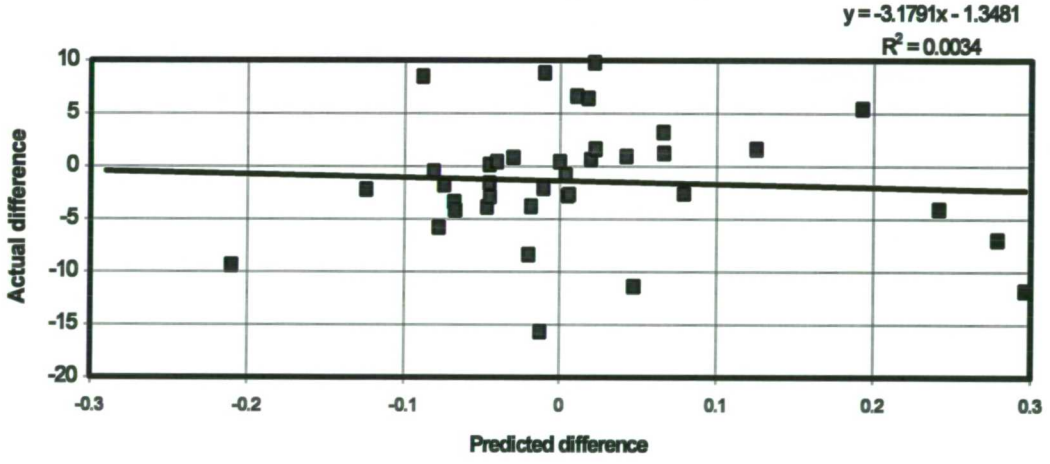


Figure 6.9

Actual differences between cumulative percentages of candidates at Grade A in 1990 and 1991, against differences predicted from changes in gender distributions



Actual differences between cumulative percentages of candidates at Grade B in 1990 and 1991, against differences predicted from changes in centre type distributions



Actual differences between cumulative percentages of candidates at Grade E in 1990 and 1991, against differences predicted from changes in centre type distributions

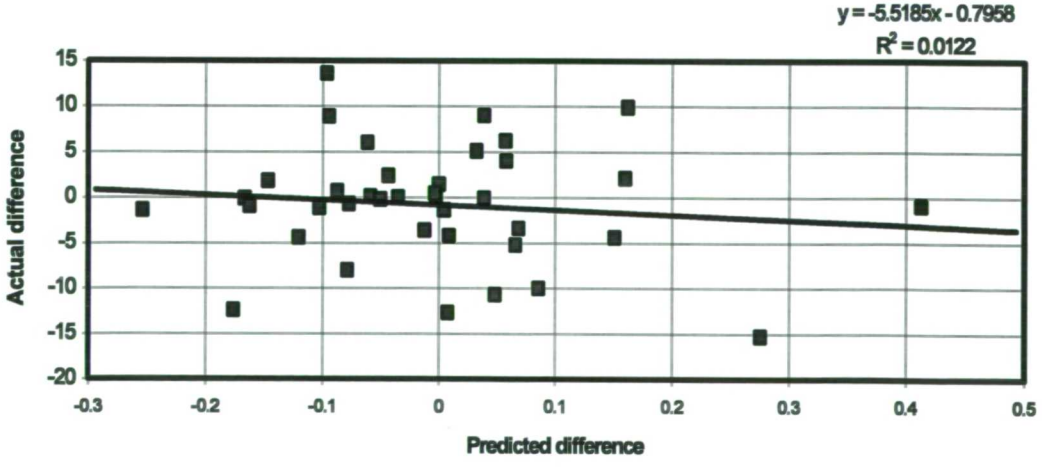
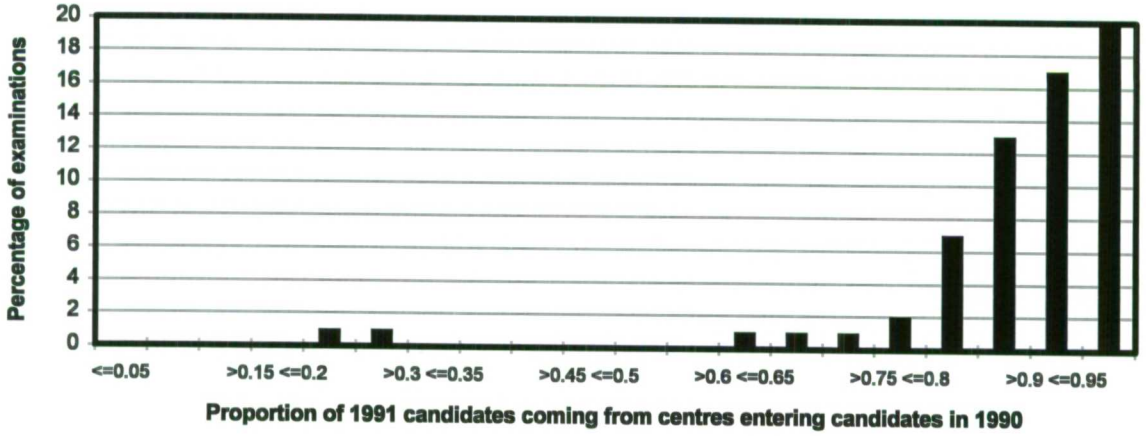


Figure 6.10

Frequency distribution of the proporton of candidates for AEB A-level examinations in 1991 who came from centres which entered candidates for the same examination in 1990



Given these data, annual variations in the relative proportions of candidates from any subgroups which differ in attainment must be small unless the proportions of these subgroups of candidates also vary in a similar way within centres. Alternatively, the same effect could be produced if the number of candidates entered by individual centres fluctuated markedly. Given the normal organisation of educational institutions, such fluctuations would be extremely surprising and do not, in fact, occur except in a few rare cases of local reorganisation of educational provision. Moreover, there is no independent evidence which suggests that one year's cohort of candidates differs in attainment or ability in a consistent way in all examination centres but differently in different subjects, nor is there any plausible extrinsic mechanism which would cause such variations, as was noted during the discussion of Explanation Number 1. Explanation Number 2 therefore seems not to be sufficient to account for the scale of changes observed in the outcomes of successive years' examinations.

6.4.3 Explanation Number 3: This year's new (missing) candidates are better (or worse) than the rest.

This explanation, which is sometimes offered by awarders when the number of candidates entering for an examination has grown or shrunk considerably, is essentially a special case of Explanation Number 2 in which the new (or missing) candidates are thought of as a subgroup of candidates with zero incidence in the previous (or current) year. As a result, the plausibility of Explanation Number 3 as a sufficient explanation for the observed changes in grade outcomes is similar to that of Explanation Number 2. Figure 6.10 implies that the proportion of candidates coming from centres which have not entered candidates before is generally low. In these circumstances, only if the entry for an examination grows or shrinks substantially as a result of many different centres making similar changes to their entry policies and entering candidates from a different range of attainment, can Explanation Number 3 account for large changes in the proportions of candidates awarded each grade. However, as Table 6.5 exemplifies, year-on-year changes in the number of candidates entering for an examination are rarely large in proportion to the existing entry and, given Figure 6.10, are the result of correspondingly small changes in the number of centres entering candidates.

However, the analysis in Appendix 6.6 implies that, unless the proportion of new (or missing) candidates is very large, and (as a group) their attainment is very different from that of previous candidates, they will not produce sizeable changes in the overall proportions of candidates awarded each grade. As an example of the scale of the effects required, consider English Syllabus III in Table 6.5 where the entry increased by almost 40% between 1990 and 1991 and there was an increase of 4.2 percentage points in the cumulative proportion of candidates awarded Grade A. Can Explanation 3 account for the change in such a case? In this subject, only 65% of the 1991 candidates came from centres which entered candidates in 1990 (about 10% of centres entering candidates in 1990 did not do so in 1991). Nonetheless, if the rise in outcomes at Grade A is explicable solely by Explanation 3, it follows that 17% of the 4874 candidates from new centres were appropriately awarded Grade A, compared with 4.2% of the remainder. This illustrates how discrepant the new group of candidates must be, even when (very rarely) they account for as much as one third of the candidates, to produce changes in outcomes like those routinely occurring in Table 6.5. (In fact, when the results of the candidates from new centres in English III in 1991 are analysed, it emerges that only 10.9% of them were awarded a Grade A and 10.3% of those from centres which entered candidates in 1990 were also awarded Grade A in 1991. Similar analyses for other examinations on other occasions [eg. Macdonald, 1992] report similarly small effects.) From this example, it can be seen that Explanation Number 3 is likely to be adequate only in rare, **and easily identifiable**, cases and it is not sufficient to account for the many significant changes in outcomes reported in Table 6.5 and Figure 6.6.

6.4.4 The relationship between the awarders' decisions and the statistics of the candidates' marks

It appears very probable, therefore, that the differences in outcomes reported in Table 6.5 are due, at least in part, to fluctuations in the standards represented by the awarders' decisions. As was noted earlier, it is not possible, in the absence of independent assessments of the candidates' achievements, to **prove** this conclusion in any particular case, nor to estimate the precise size of the discrepancies which occur. However, it is possible to establish upper bound estimates for the movements in grade boundaries which would have been required to set standards in 1991 which were comparable to those in 1990. If it is assumed that the

attainments of the 1991 candidates were distributed identically to those of the 1990 candidates, then changes in the distributions of marks between the two years can be interpreted as indicating changes in the difficulty of the question papers and/or changes in the severity of the marking process. Since the purpose of awarding is to make adjustments to grade boundaries which compensate for such changes, estimates of the positions of the 1991 grade boundaries can then be obtained by scaling the grade boundaries used in 1990 in accordance with the means and standard deviations of the 1990 and 1991 mark scales, as follows:

$$B'_x = \frac{(B_x - m_{90})}{s_{90}} \bullet s_{91} + m_{91}$$

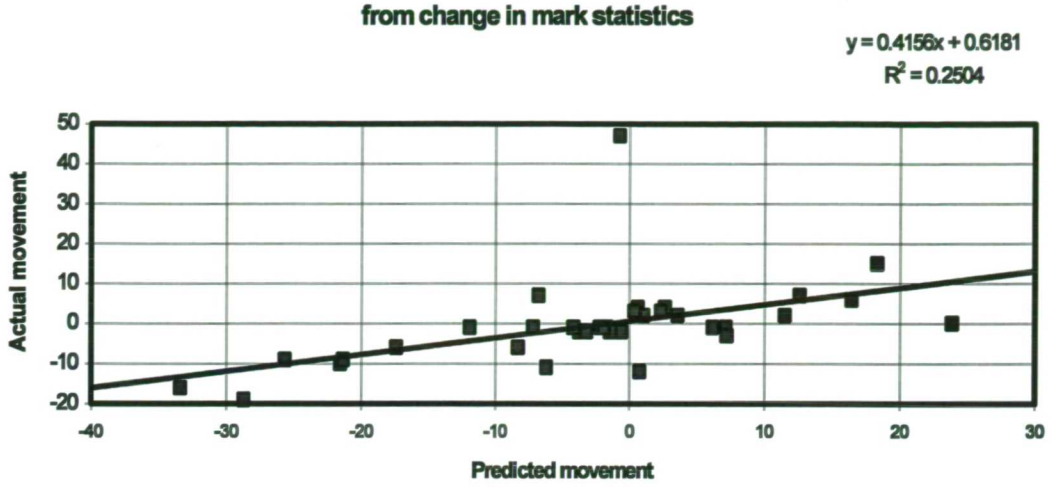
where B'_x is the new position of the boundary B_x for Grade x
m_y is the mean score in year y
and s_y is the standard deviation of scores in year y

The results of doing this have been compared with the positions of the grade boundaries in 1991, producing Figure 6.11 which plots the actual movements of the grade boundaries against the movements predicted on the basis of the changes in the mark statistics. (Two of the examinations in Table 6.5 have been excluded because their maximum mark changed substantially between 1990 and 1991, producing misleadingly extreme movements.) It can be seen that there is a fairly strong relationship between the predicted and actual grade boundary movements but that, on average, the size of the actual movements is about 0.4 of the size of the predicted ones. From this analysis, it appears that the awarders correctly identified the direction of the changes required but, given the present assumption that the candidates were of comparable achievement in the two years, failed to take sufficient account of the change in difficulty of the examination papers and/or their marking.

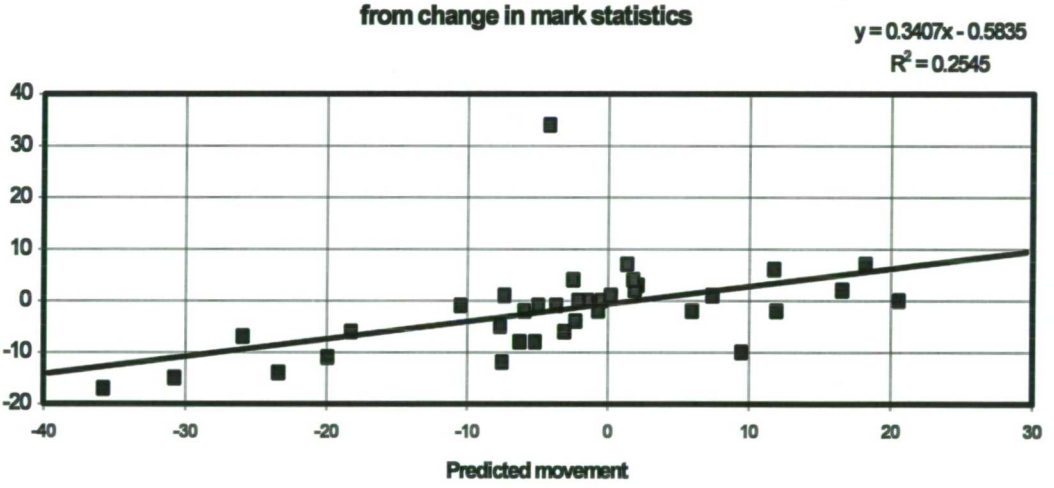
Although the assumption upon which the present analysis is based is exactly that, an assumption, it is important to note that, in 81 (77%) of cases, the awarders' decisions moved the boundaries in the **direction** implied by the mark statistics, if not to the **extent**. Thus, in these cases, the awarders' judgements confirm that, to some extent at least, the mark statistics reflect changes in the difficulty of the examinations. However, there is no reason to

Figure 6.11

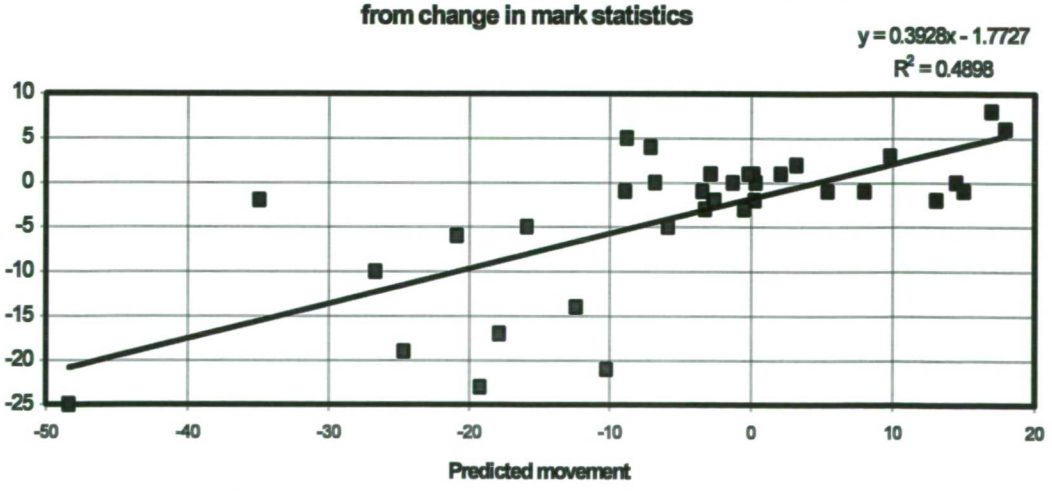
Actual movement of Grade A boundary in 1991, against movement predicted from change in mark statistics



Actual movement of Grade B boundary in 1991, against movement predicted from change in mark statistics



Actual movement of Grade E boundary in 1991, against movement predicted from change in mark statistics



believe that any change there might be in the attainment of the candidates from one year to the next is not independent of any change in the difficulty of the examination which occurs. Therefore, if the reasonable assumption is made that the candidates are equally likely to be slightly better or slightly worse from one year to the next, the actual movement of the grade boundaries should be less than that predicted from the change in mark statistics in 50% of cases and greater than the predicted change in the remaining 50%. However, of the 81 cases where the mark statistics and awarders agree on the direction of the move, the actual move is less than the predicted move in 57. Using the binomial distribution, the two-tailed probability of this (or a more extreme value) occurring by chance is easily shown to be less than 0.001. This strongly suggests that, for some of these examinations at least, the awarders' judgements took insufficient account of changes in the difficulty of the examination papers and/or their marking. This conclusion is consistent with Good and Cresswell's (1988a and 1988b) experimental result that awarders tend towards relative severity when setting grade boundaries on harder papers within differentiated paper examinations. Thus, it seems reasonable to conclude that, as the evidence in Section 6.3 suggests, at least some of the Phase 2 meetings did not take sufficient account of the context of candidates' performances and were using, to some extent at least, the inappropriate *Script as Artefact* awarding strategy.

6.5 THE PHASE 3 OBSERVATIONS OF AWARDING MEETINGS IN 1993

The weight of evidence in the preceding section implies that some of the observed changes between 1990 and 1991 in the outcomes of the examinations were artefacts of the awarding process rather than the consequences of changes in the attainment of the candidates. Thus, in so far as the existence of such artefacts implies that, in some examinations, the previous years' candidates would have received different grades if they had taken the present year's examination, a failure to maintain comparable grading standards between the two years is implied for at least some of the subjects. The analysis of the behaviour and concerns of the awarders which was reported in Chapter 5 and Sections 6.2 and 6.3, indicates some of the possible underlying reasons why such changes might occur.

As reported in Chapter 5, a number of amendments were therefore made to the awarding procedures, and to the briefing given to awarders, following the 1991 examinations. These amendments were described in detail in Section 5.6.1. In Summer 1993, further systematic observations were made of A-level awarding meetings to investigate the effects of the changes.

6.5.1 The meetings observed in Phase 3

Practical difficulties related to the timing of meetings and the need for the author to attend certain other meetings as a participant meant that it was not possible for awards in exactly the same subjects to be observed in Phase 3 as had been observed in Phase 2. The meetings observed in Phase 3, in the order in which they occurred, were as follows:

1. Economics,
2. Physics,
3. Mathematics,
4. English Language and Literature,
5. Communication Studies.

6.5.2 The coding scheme used in Phase 3

For Phase 3, it was not possible, for the same practical reasons, for the author to carry out the observations. Two observers were therefore recruited to do this. They were briefed on the meaning of the categories to be used and equipped with forms on which to record their observations (a copy of the form is to be found in Appendix 6.7). To simplify the task for these less informed observers, one minor change was made to the observational categories: the distinction between procedural and methodological remarks was dropped. The slightly reduced category system used for Phase 3 was therefore as follows:

Positive affect/social	Negative affect
Gives methodological or procedural guidance	Seeks methodological or procedural guidance
Gives evaluative criterion	Seeks evaluative criterion
Gives overall judgement (cands as a group)	Seeks overall judgement
Gives evaluation of a particular script	Seeks evaluation of a particular scriptAsks
Gives statistical opinion or information	Seeks statistical opinion or information
Gives opinion or information concerning the paper	Seeks opinion or information concerning the paper
Makes other relevant point	Seeks other relevant information
Suggests boundary mark	Asks for boundary suggestion

The observers were asked to encode the remarks of the Chairs, awarders and Board officers separately. As with Phase 2, although the seeking and giving of particular types of opinion and information were separately encoded, the data for seeking and giving have been combined in the analyses reported in this section. Appendix 6.8 contains the complete summarised raw data from the Phase 3 observations, aggregated across all papers and grade boundaries within each subject.

6.5.2.1 Reliability of the Phase 3 observations

The use of two observers, working independently on the same meetings, enabled the reliability of the Phase 3 coding to be assessed. In addition, the author encoded the awarders' discourse concerning one paper (in the Economics meeting) alongside the two observers so that the agreement between the observers and the author could be assessed. The data from the observations of these three observers on this paper are contained in Appendix 6.9 and the differences between them can be seen to be small. The statistical significance of the differences was tested using χ^2 by treating the data for each type of participant (chairs, officers and awarders) as a separate contingency table of observers versus types of remark. The differences between the two observers (see Appendix 6.8) were tested in the same way for the other meetings which they both observed (for practical reasons, the observers were not both able to observe the Communication Studies meeting). The results of all these analyses are shown in Tables 6.6 and 6.7

Table 6.6
The reliability of the Phase 3 observations; significance of the differences among the observers and author (Economics Paper 3)

	Chairs	Awarders	Officers
χ^2	5.55	11.01	8.21
degrees of freedom*	12	14	4
significance level	0.94	0.69	0.99

* Categories excluded from analysis if expected value for any observer < 1 (see Everitt, 1979).

Table 6.7
The reliability of the Phase 3 observations;
significance of the differences between the two observers

	Physics			Mathematics		
	Chairs	Awarders	Officers	Chairs	Awarders	Officers
χ^2	6.03	9.20	1.33	1.05	7.37	3.92
degrees of freedom*	7	7	1	7	7	6
significance level	0.54	0.24	0.25	0.99	0.39	0.69
	Economics			English		
	Chairs	Awarders	Officers	Chairs	Awarders	Officers
χ^2	1.20	2.54	1.20	4.42	12.43	7.42
degrees of freedom*	7	6	5	6	7	5
significance level	0.99	0.86	0.94	0.62	0.09	0.19

* Categories excluded from analysis if expected value for either observer < 1 (see Everitt, 1979).

The reliability of the Phase 3 observations was evidently high and, since none of the differences between the observers were statistically significant beyond the 0.05 level, the data from them have been pooled for the remainder of the analyses reported here.

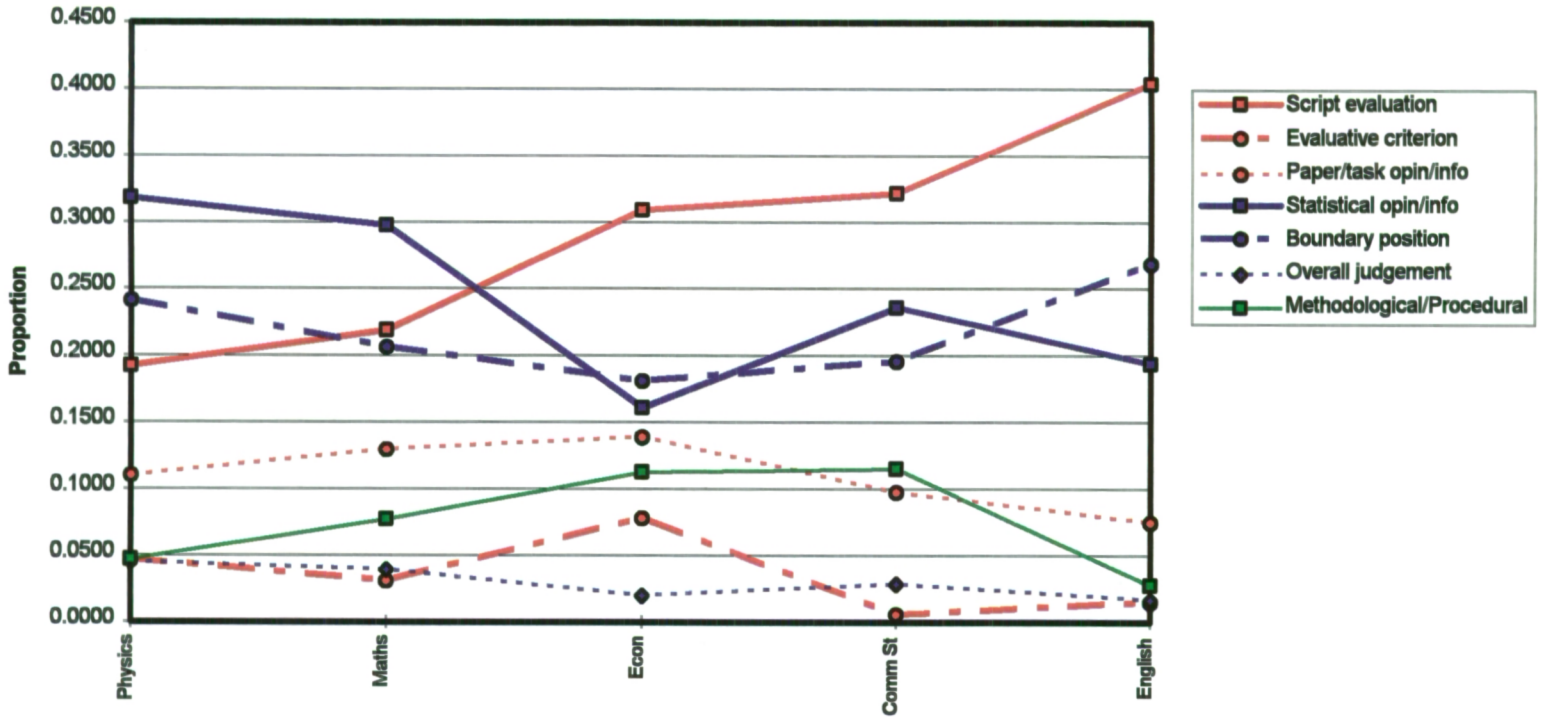
6.5.3 The focus of the awarders' discourse in Phase 3

Figure 6.12 shows the proportion of participants' remarks falling in each of the observational categories which is directly relevant to the task of awarding. (Affective/social combinations have been omitted from Figure 6.12; their frequency is reported in Appendix 6.8 and Figures 6.13 to 6.15)

A comparison of Figure 6.12 and Figure 6.1 shows that the same broadly inverse relationship still existed in 1993 between the proportion of evaluations of individual scripts and the proportion of references to statistical data. However, in 1993 those meetings in which statistical commentary predominated (for example, Physics) involved more explicit evaluation of individual scripts than in 1991, whereas those in which individual evaluations predominated (for example English) involved more reference to statistical data than in 1991. It appears, therefore, that the change in procedures between 1991 and 1993 affected the balance of

Figure 6.12

Proportion of observed participants' remarks in each category (excluding affective/social) - Phase 3



evidence considered by the awarders, leading, in the meetings observed, to a more even-handed approach to statistical and qualitative data. In connection with this, it is worth noting the increase between Phases 2 and 3 in the proportion of remarks about the position of the grade boundaries in Communication Studies and English. This is also consistent with the 1993 requirement explicitly to consider both main types of evidence when reaching boundary decisions, from which more discussion of the precise position of the boundaries is to be expected.

There are four other points particularly worth noting from the comparison of Figures 6.12 and 6.1. First, there was considerably less discussion of methodological and procedural matters in Phase 3. This presumably reflects the more extensive briefing given to the awarders by the 1993 procedure papers. Second, the amount of comment in the meetings involving overall judgements of the candidates as a group was reduced. Again, the dubious validity and relevance of such discussion to the main task of the meeting (see Chapter 5) was pointed out in the 1993 awarding procedure files. On the other hand, despite the new procedures, there was no more contextualising discussion of the question papers than in 1991. Third, the proportion of explicit discussion about evaluative criteria was, again, very low. Finally, the very high proportion of statistical remarks in the 1991 Mathematics meeting was not repeated in 1993 suggesting, as surmised earlier, that this reflected the particular difficulties encountered with one very hard paper in the 1991 Mathematics meeting (see Chapter 5).

The χ^2 analysis of residuals which was carried out on the frequencies of the different types of remarks for Phase 2 was repeated for the Phase 3 data, with the results shown in Table 6.8, below. The differences between the meetings were again highly statistically significant (χ^2 = 167 with 24 degrees of freedom).

Direct comparisons between Tables 6.8 and 6.1 are not possible because they represent different groups of subjects, albeit with four in common. However, the same phenomenon of significantly more emphasis on statistical data , and significantly less emphasis on script evaluation in Physics and Mathematics is again present, with the reverse situation occurring in English.

Table 6.8
Adjusted residuals for frequencies of remarks in each category (excluding affective/social) - Phase 3.

Nature of remark	Physics	Maths	Economics	Communication studies	English
Script evaluation	-3.50	-5.77	0.86	0.75	**7.98**
Evaluative criterion	1.21	-0.56	**4.15**	-1.48	-3.20
Paper/task opin/info	-0.03	**2.33**	1.50	-0.39	-3.51
Statistical opin/info	**2.66**	**4.29**	-3.46	-0.30	-3.96
Boundary position	0.53	-1.95	-1.88	-0.73	**3.03**
Overall judgement	1.42	**2.06**	-1.01	-0.09	-2.47
Methodological/ Procedural	-1.04	**2.44**	**3.54**	**2.10**	-4.37

Note - **bold** and underlined text respectively indicate those positive and negative values which are statistically significant (< 0.05).

6.5.4 The different roles of the participants in the Phase 3 meetings

In this section, data on the differences between the contributions made to the Phase 3 meetings by the chairs, awarders and officers are reported. Table 6.9 reports the results of a χ^2 analysis of residuals on the frequencies of each category of contribution when the data from all the Phase 3 meetings are combined. Overall, the differences between the frequencies for nature of remark by role were highly statistically significant (χ^2 = 393 with 14 degrees of freedom).

Comparison of Tables 6.9 and 6.2 shows a very similar pattern of significant differences between the contributions of the participants in Phases 2 and 3. As before, the awarders concern themselves primarily with the more qualitative aspects of the awarding process, referring significantly more often than the chairs and officers to evaluative criteria, the papers and tasks involved and the overall quality of the candidates. The awarders also significantly more often mention the quality of individual scripts. The officers make significantly more references to the statistical data and to procedural and methodological matters. The chairs of the meetings make significantly more affective/social contributions to the meetings than the other participants and also mention the positions of the boundaries significantly more often than the other participants. These data are consistent with the roles of the participants in the meeting, as defined by the Board's new procedural documentation (see Appendix 5.2).

Table 6.9
Adjusted residuals for frequencies of remarks in each category by role - all Phase 3 meetings.

Nature of remark	Chairs	Awarders	Officers
Script evaluation	-1.24	**7.67**	-9.81
Evaluative criterion	-1.60	**2.61**	-1.64
Paper/task opin/info	-5.29	**7.09**	-3.13
Statistical opin/info	-2.05	-7.62	**14.44**
Boundary position	**2.36**	-2.85	0.94
Overall judgement	-1.48	**2.63**	-1.85
Methodological/ Procedural	1.91	-4.42	**3.94**
Affective/Social	**7.13**	-3.37	-5.12

Note - **bold** and underlined text respectively indicate those positive and negative values which are statistically significant (< 0.05).

However, as before, these aggregate data inevitably disguise differences between the meetings. Table 6.10 shows the proportions of the contributions to each meeting made by chairs, awarders and officers. The differences across the meetings in the frequencies of contributions by those with different roles are again highly statistically significant ($\chi^2 = 92$ with 8 degrees of freedom). Analysis of standardized residuals shows the Physics and Economics chairs to have made significantly more contributions, and the Mathematics chair significantly fewer, than the English and Communication Studies chairs. The Officers in the physics meeting made significantly fewer contributions than their peers in the other meetings.

Table 6.10
Proportions of contributions to each Phase 3 meeting by chairs, awarders and officers

	Physics	Maths	Economics	Commun. Studies	English
Chairs	0.48	0.24	0.47	0.41	0.32
Awarders	0.45	0.62	0.40	0.45	0.49
Officers	0.07	0.14	0.14	0.14	0.12

Figure 6.13

Proportions of remarks in each category from chairs of Phase 3 meetings

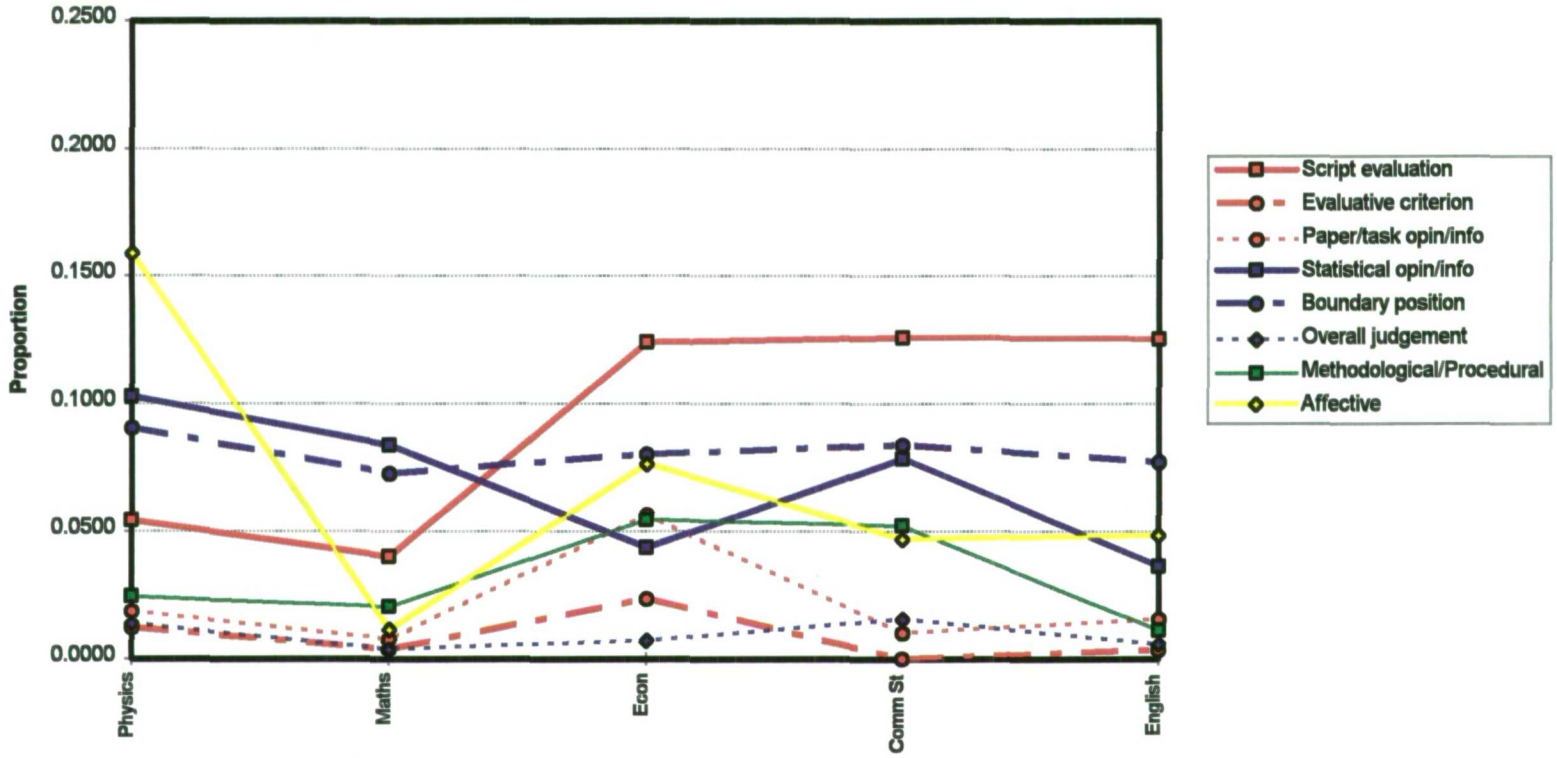


Figure 6.14

Proportions of remarks in each category from awarders in Phase 3 meetings

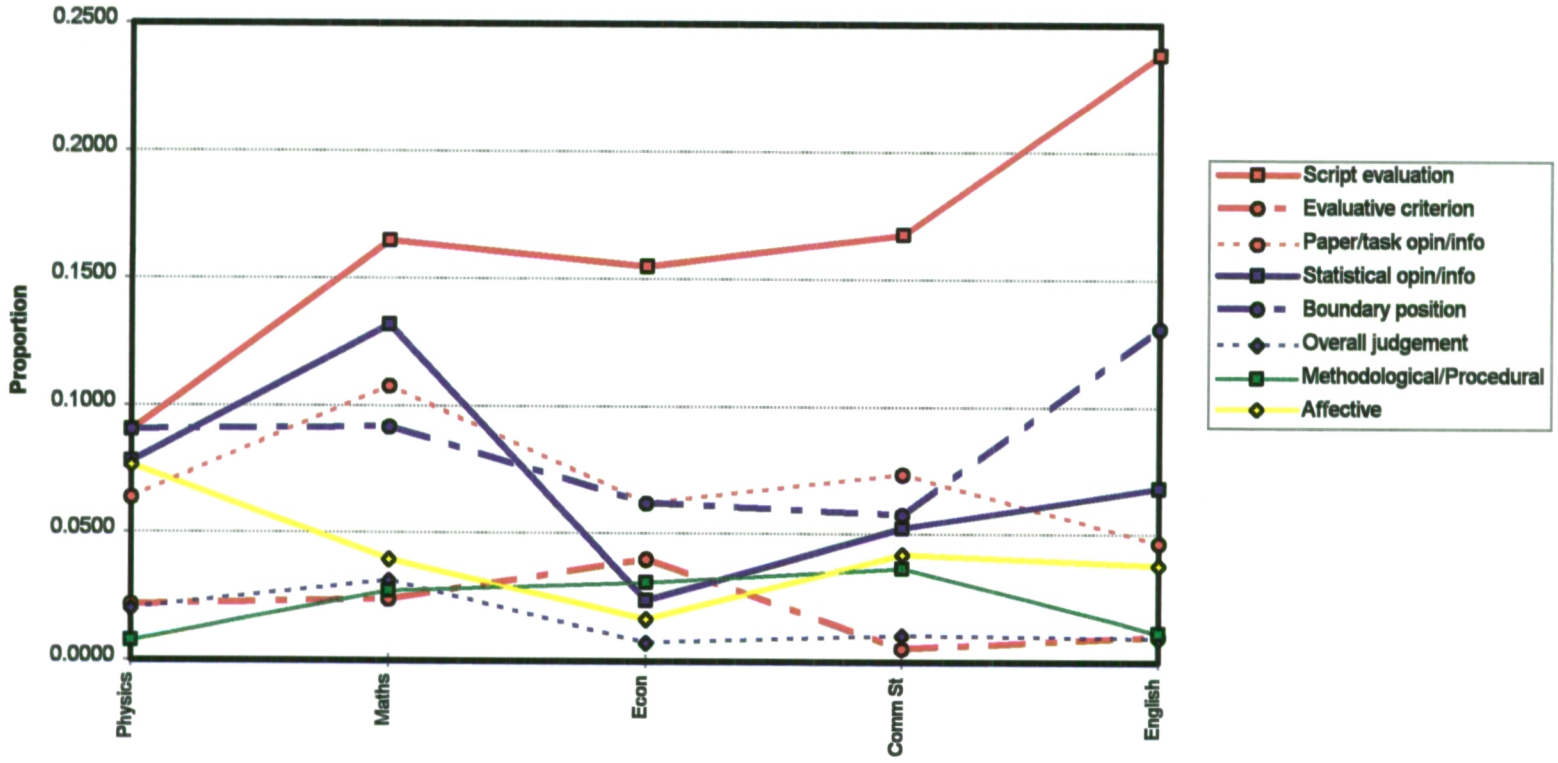
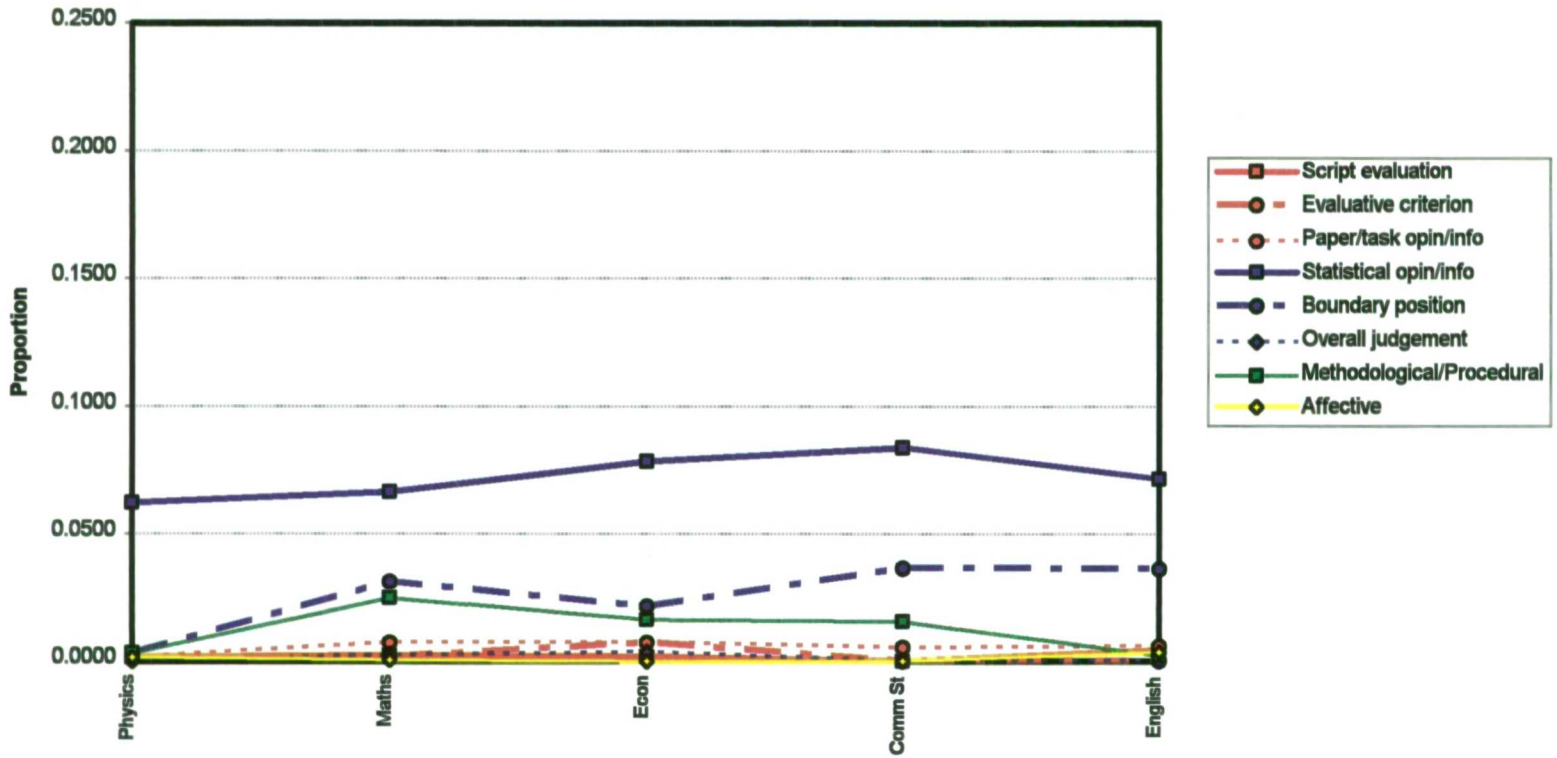


Figure 6.15

Proportions of remarks in each category from officers in Phase 3 meetings



Figures 6.13 to 6.15 show, separately for the chairs, awarders and officers respectively, the proportions of remarks in each category. From these graphs, several points emerge. First, although, compared with the awarders, a greater proportion of the chairs' contributions are generally affective and social or concerned with statistics, there are also considerable differences between the contributions of the chairs in different meetings. As with Phase 2, there is a marked similarity between the emphases on qualitative and statistical data in the remarks made by each chair and the emphases for each meeting as a whole (compare Figures 6.13 and 6.12). This reflects, to some extent at least, the influence of the chairs on the tenor of the meetings. The chairs appear to be partly responsible for increasing the emphasis on statistical data in each meeting as a whole since, in all the meetings (even Physics and Mathematics), the awarders' most frequent contribution to the discussion still takes the form of expressing evaluative judgements about individual scripts (Figure 6.14). The awarders' emphases on script evaluation and statistical data still vary between the meetings in a way which is similar to, but more extreme than, the variations in emphasis of the chairs. Finally, the officers' contributions to the meetings are now very similar across subjects and consist predominantly of the provision of statistical information and opinion.

In summary, the changes in the Board's procedures between 1991 and 1993 appear to have had a number of small but positive effects in terms of the delineation of the roles of the participants in the awarding meetings. The awarders in all subjects now focus primarily on evaluative judgements, the chairs maintain a balance between the qualitative and quantitative evidence and the officers now ensure that statistical data are presented even when the awarders have a clear preference for more qualitative evidence. Overall, there was a more even-handed approach to statistical and qualitative evidence, though differences between subjects in this respect still remained in the subjects observed in Phase 3.

6.6 THE OUTCOMES OF AWARDING MEETINGS UNDER THE NEW PROCEDURES

The question that remains is how the changes in behaviour of the participants in the awarding meetings, which were engineered between Phases 2 and 3 of the observational work, affect the outcomes of the process. In particular, are annual changes in grade distributions still

occurring on the scale and with the frequency reported in Section 6.4? This section addresses this question by comparing the 1993 and 1994 outcomes for the same examinations as were analysed in Section 6.4. The years 1993 and 1994 were chosen for this analysis because in 1992 and 1993 the effects of another change in awarding procedures was working its way through the statistical data. Specifically, as reported in Chapter 5, the board adopted the percentile method for aggregating component grade boundary decisions (see Chapter 4) during this period. This greatly complicates comparisons of outcomes between 1991, 1992 and 1993. However, the new aggregation method was established by 1993 so that the comparison of 1993 outcomes with those of 1994 is comparable with the comparison between 1990 and 1991. There is no reason to expect that the use of the percentile aggregation method, once established, has any impact on the stability of the examination outcomes from one year to the next.

Table 6.11 shows the changes in outcome between 1993 and 1994 for the same examinations as were considered in Table 6.5. It is immediately apparent that the scale of the changes is very much reduced and Figure 6.16 shows the distribution of z statistics obtained by testing these differences in the way described in Section 6.4.1. It is comparable, therefore, with Figure 6.6. Again, the difference is striking. The changes in outcomes between 1993 and 1994 were only slightly greater than would be expected as the result of chance variations between cohorts of candidates.

One feature of Figure 6.16 that is particularly interesting is the tendency towards improved grades in 1994, compared with 1993. A further year's data would be needed to confirm whether this is a feature of the new awarding procedures but it seems quite possible that it is. It was noted in Chapter 5 that awarding committees tend to be slightly more sanguine about increasing results than about decreasing them. The greater stability between years shown in Table 6.11 is almost certainly due to greater attention being paid by the awarders to statistical evidence because this evidence alerts the awarders at an early stage in their deliberations to possible large changes in candidates' results which they can then avoid. If they are slightly less concerned about increases in candidates' results than about decreases (see Chapter 5),

Figure 6.16

Distribution of z statistics for differences between outcomes in 1993 and 1994 for each key grade

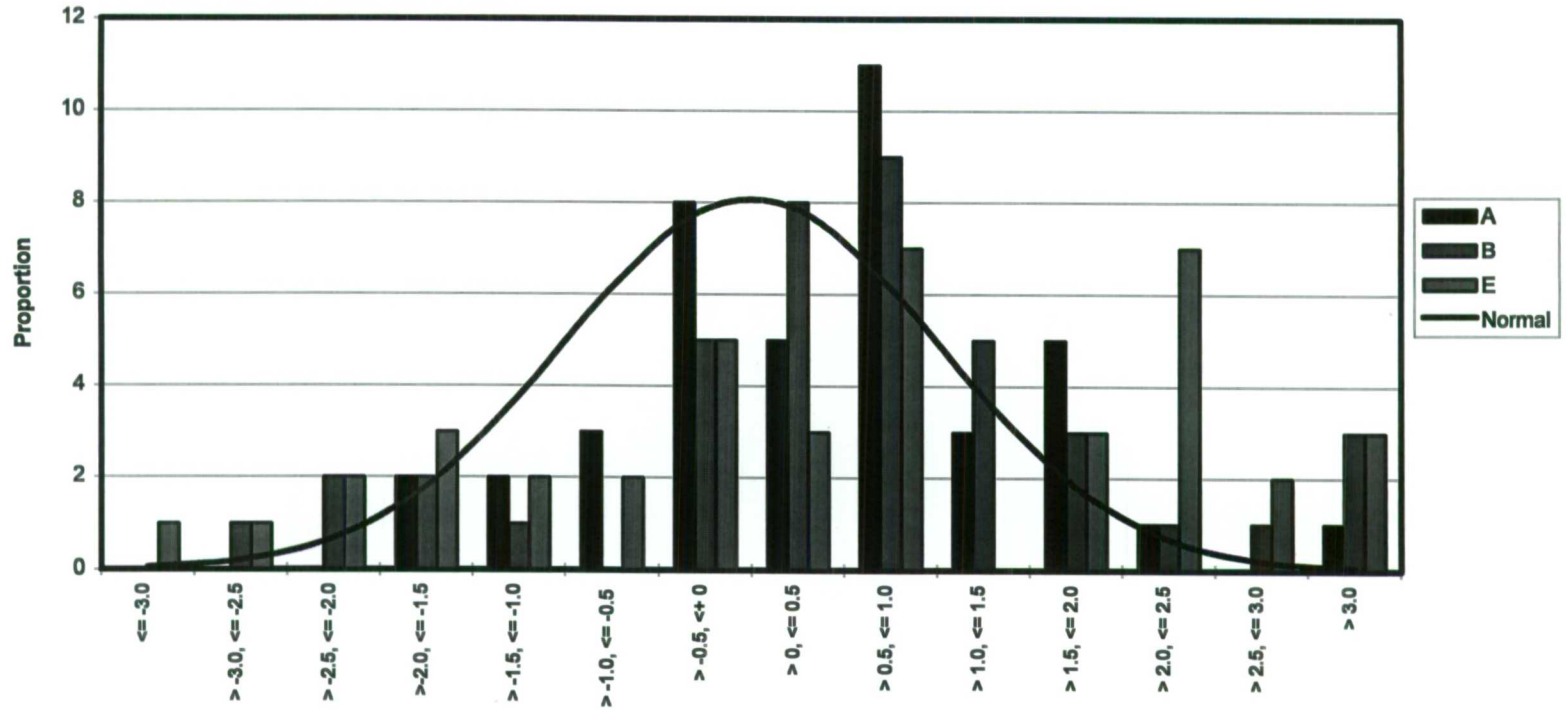
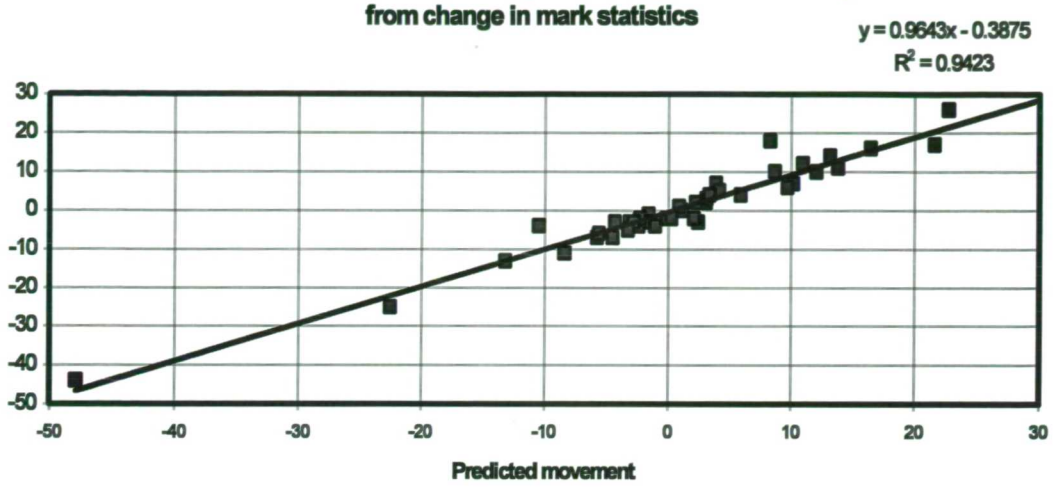


Table 6.11
Comparisons between the outcomes in 1993 and 1994 for A-level examinations with more than 500 candidates in 1991

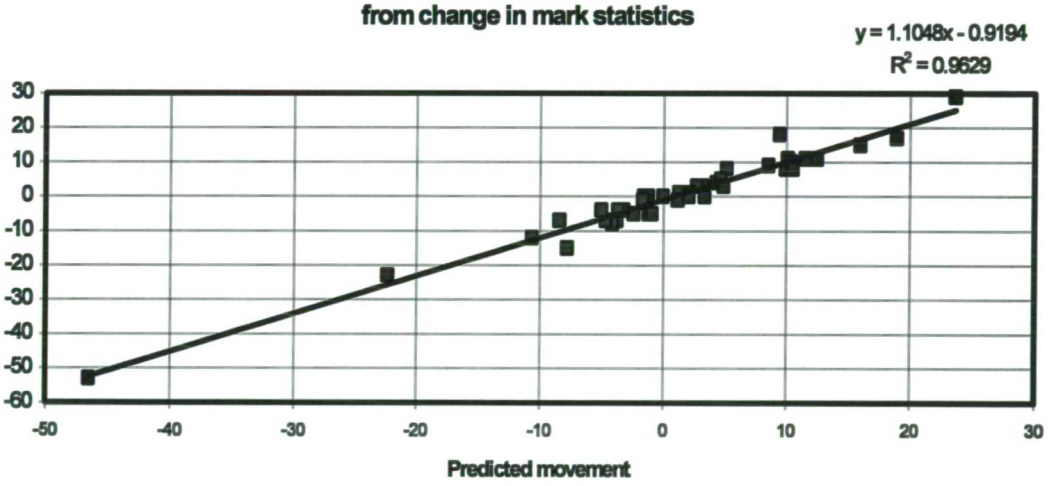
Subject	Number of cands in 1993	Number of cands in 1994	Change in cum % at Grade A	Change in cum % at Grade B	Change in cum % at Grade E
Accounting	4630	4105	1.2	0.3	-1.6
Applied Mathematics	836	773	-4.6	-4.4	-0.1
Biology I	6333	7016	1.1	0.9	-0.3
Biology II	2350	2654	0.4	1.4	0.7
Business Studies	14863	16018	1.2	1.5	-0.1
Chemistry	3476	3262	0.9	0.4	-0.5
Communication Studies	5303	5571	-0.9	-2	0.4
Computing	3988	4063	1.2	-0.1	-4.4
Constitutional Law	558	496	-0.2	0.5	2.4
Economic & Social History	1135	1012	-0.6	-0.4	-2
Economics	9134	7331	0.9	0.6	2.7
English I	14540	14774	-0.3	1.6	2.1
English II	2906	2367	-0.4	0.7	2.7
English III	19341	19698	-0.7	0.3	0.2
Environmental Science	1463	1699	0.4	1.1	3.2
French	5731	5345	-0.2	-2.2	-1.5
General Studies	1901	2305	-0.3	-0.2	1.6
Geography	3616	3745	0.2	0	-2
German	2291	2187	0.2	1.1	1.8
Government & Politics	1089	918	1.2	0.7	0.7
History	2433	2515	1.2	2.1	2.2
History (Alternative)	1497	1638	-0.1	0.2	-0.9
History of Art	830	811	-1.1	-2.8	-2.7
Human Biology	3624	3956	0.6	0.8	2.5
Law	4421	4592	-0.4	0.3	2.2
Philosophy	920	967	1	2.4	4.3
Photography	1575	1578	0.9	2.3	2.6
Physical Education	2499	3621	0.4	0.8	0.2
Physics	5736	5298	0.7	2.0	4.3
Psychology	15452	18537	0.3	1.2	0.4
Pure & Applied Mathematics	5195	4049	1	0.3	2
Pure Mathematics	1944	1682	2.4	3	-0.2
Pure Mathematics & Statistics	3944	3192	0.9	0.8	3.3
Sociology I	18087	17036	-0.1	0.7	-0.8
Sociology II	7741	9372	0	-2	-1.7
Spanish	1067	1054	-0.1	0.2	-0.3
Sport Studies	1616	2015	0.3	3.7	1
Statistics	1089	1062	1	1.3	0.5
Theatre Studies	7088	7789	0.5	-1	-1.2

Figure 6.17

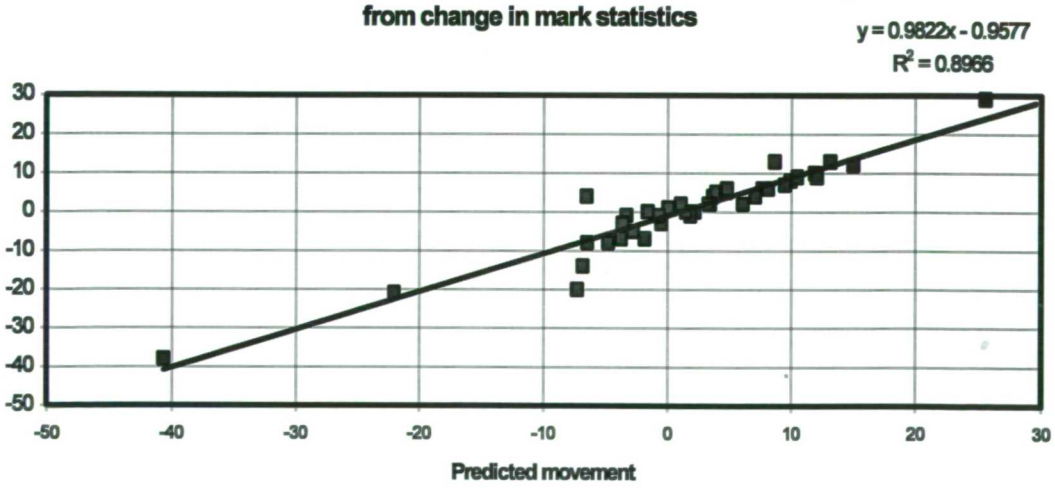
Actual movement of Grade A boundary in 1994, against movement predicted from change in mark statistics



Actual movement of Grade B boundary in 1994, against movement predicted from change in mark statistics



Actual movement of Grade E boundary in 1994, against movement predicted from change in mark statistics



then the bias evident in Figure 6.16 might be expected once the awarders are well informed about the statistical consequences of their decisions. The alternative hypothesis is, of course, that there was a small improvement in the attainment of the candidates for many of the examinations analysed in 1994 but the probability of this explanation is low for the reasons given in Section 6.4.

The analysis reported in Section 6.4.4 was repeated for the 1993/1994 data with the results shown in Figure 6.17. It can be seen that, unlike 1991, in 1994 the awarders moved the boundaries in most, though not all, examinations in the direction **and, on average, to the extent** implied by the changes in mark statistics between the two years. Of the 95 (out of 105) cases where they moved the boundaries in the direction implied by the mark statistics, 47 such movements were smaller than predicted and 48 greater than predicted; as close to the theoretical 50:50 split as possible (see Section 6.4.4).

It seems probable that the selection of scripts for review on the basis of the relationship between the two years' mark scales (see Section 5.6.1 and Appendix 5.2) had a considerable effect upon the final positions of the grade boundaries. This would be consistent with the literature on the psychology of judgement. For example, Eiser (1990) quotes evidence to suggest that, in terms of Kahneman and Miller's (1986) norm theory, stimuli in judgement tasks are compared with 'local norms' rather than with the whole range of possible positions along the dimension being judged.

6.7 IN CONCLUSION

The amended awarding procedures observed in Phase 3 appeared to have produced a number of improvements in awarders' practice. There were still subject specific differences in the emphasis given to statistical and qualitative data but all the meetings considered both sorts of evidence. There was less uncertainty about procedures and methodology, the roles of the participants were more clearly delineated and there was less irrelevant speculation about the overall attainment of the candidates.

The amount of explicit contextualising discussion of the difficulty of the question papers remained low but the use of the mark statistics to determine which scripts were scrutinised in the awarding meeting appeared to contextualise the process. Certainly, the new procedures, including the provision of more statistical data at an earlier stage in the meeting, produced greater stability in the examination outcomes from year to year so that, overall, the outcomes in 1994 are much more consistent with those in 1993 than were the 1991 outcomes with those in 1990. Compared with 1993, there may have been a slight tendency towards lenience in 1994 but the data reported in this section are consistent with the hypothesis that, for the examinations analysed, the amended awarding procedures led to comparable standards being set in 1993 and 1994.

On the other hand, it is possible there were systematic changes in candidates' attainment in some subjects between 1993 and 1994 which were not reflected in the examination outcomes under the new procedures and that the consistency between the 1993 and 1994 outcomes is, in fact, too great. There are no independent data available to test this possibility, the implications of which will be considered in more detail in Chapter 9.

CHAPTER 7

GRADE AGGREGATION - A CASE STUDY

"Two times two is twenty two
Four times four is forty four"
When Numbers get Serious - Paul Simon

7.1 INTRODUCTION

In this chapter, a case study of the use of grade aggregation to derive subject grades directly from component grades is reported. Grade aggregation was described in general terms in Chapter 4 (Section 4.2.1) where it was proposed as a possible alternative to conventional mark aggregation. It avoids the necessity for combining component grade boundary judgements, the problematic nature of which was discussed in Section 4.2.2 but at the cost of some theoretical reduction in the reliability of the grades awarded for the examination as a whole (Section 4.2.1). The purpose of the work reported in this chapter was to explore some of the practical consequences of grade aggregation from the perspective of the theoretical concerns set out in Chapter 4, particularly the effects of the theoretical reduction in reliability. An additional feature of the work is the development of new techniques for evaluating the component weightings represented by any particular set of grade aggregation rules. No technique for doing this has been found in the published literature.

The particular practical use of grade aggregation explored in this chapter was in a modular A-level examination in Chemistry. The examination in question was the first of a larger suite of modular examinations covering a number of subjects and known as the *Wessex Project* examinations after the modular curriculum project for which they were provided. According to Wilmut (1990), the principal reasons why these examinations used grade aggregation to derive final subject grades included those given in Chapter 4 and were:

> to enable candidates to interpret their module grades in terms of their likely final subject grade so that they could choose when to cash-in their accumulating module credits and when to re-take a module or take an alternative one;
>
> to avoid apparent anomalies between the profile of module grades and the subject grade;

to place the results from all modules onto a common scale (the grade scale) and thus enable different combinations of modules to be treated interchangeably within the subject grading process.

7.2 THE SCHOOLS AND CANDIDATES

The students providing the data reported in this chapter were those involved in the second year (1990) of the pilot operation of the Wessex Chemistry examination. In that year, 153 candidates were entered for certification by 8 schools (112 candidates) and 3 Further Education colleges (41 candidates). Of these 153 candidates, data from 147 were available for analysis and formed the basis for the work reported in this chapter. (There was nothing exceptional about the remaining 6 candidates other than their absence, for purely logistical reasons, from the board's main database when the data were extracted.) The schools involved were comprehensive schools distinguished from others of their type mainly by their involvement in the delivery of a new modular curriculum through the Wessex Project. Since the project operated in central South West England, all the pilot schools and colleges came from that geographical region.

The schools, colleges and students providing the data reported in this chapter are not, therefore, necessarily representative of all the schools, colleges and candidates who normally enter for A-level examinations in Chemistry or, indeed, in other subjects. However, they provide a distribution of attainment in A-level Chemistry sufficient for the purpose of exploring the operation of the aggregation rules. Research questions such as the comparability of grade standards between the Wessex examination and other A-level Chemistry examinations, which might have required a more representative sample, are not addressed in this chapter.

7.3 THE EXAMINATION AND ITS OPERATIONAL AWARDING PROCEDURE

As in all the Wessex modular examinations, the candidates for Wessex Chemistry took a conventional set of examination components known as the *core examination* and, during their course, completed four *modules* chosen from a much larger set. According to the syllabus documents, the core examination was intended to contribute 60% towards the final results and the modules, aggregated together, to contribute the remaining 40%.

The core examination had three conventional components as follows: Paper 1 which was intended to provide 25 of the 60% core weighting; Paper 2 which was intended to provide a further 25 of the 60% core weighting and the assessment of practical work, assessed by the teachers and moderated by the Board, providing the remaining 10 of the 60% core weighting. The core examination was conventionally awarded in the way described in Chapter 5 except that, although the awarders made their judgements in terms of conventional A-level grades, the results were reported as Levels 1 to 6 (plus an ungraded category U) rather than Grades A to E (plus N and U). The functioning of the core examination will not be discussed here except to note that it was satisfactory. Further details concerning the papers and their technical characteristics can be found in Cresswell (1990).

The modules were assessed as coursework by the candidates' teachers and externally moderated by the Board. They were intended to be delivered by supported self-study and each candidate produced, for each module, a folder of written work. The assessment of these folders required the teachers to assign each of them, via an initial marking process, to one of a set of four levels (plus an ungraded category, U) using a single set of generic criteria which applied to all the modules. The levels for each candidate's modules were then added together to form a module sum on a scale ranging from 4 (best) to 16 (worst). Further details of the module assessment and moderation processes can be found in Cresswell (1990).

7.3.1 The aggregation rules

The core examination levels were then combined with the sum of the module levels by means of the aggregation rules shown in Table 7.1. These had been drawn up, independently of the award meeting, by the development team for Wessex A-levels and applied to all the Wessex examinations.

In Step 4 of the core examination awarding process (which was done in a separately constituted *verification meeting*) the appropriateness of the outcomes of the aggregation rules was judged by means of an holistic inspection of individual candidates' complete work: core examination papers and four module folders. The possibility existed at this stage of modifying

the core examination level boundaries to achieve, via the predetermined combination rules, final grades which were judged holistically appropriate by the awarders. The Board's procedure papers for the awarding of Wessex project examinations form Appendix 7.1.

Table 7.1
Full aggregation rules for Wessex modular A-level examinations

Module Sum		Core Level						
		U	6	5	4	3	2	1
	4	U	D	D	C	B	A	A
	5	U	D	D	C	B	A	A
	6	U	E	D	C	B	B	A
	7	U	E	D	C	C	B	A
	8	U	E	D	D	C	B	A
	9	U	E	E	D	C	B	A
	10	U	E	E	D	C	B	B
	3+U or 11	U	N	E	D	C	C	B
	12	U	N	E	D	D	C	B
	4+U or 13	U	N	E	E	D	C	B
	14	U	N	N	E	D	C	C
	5+U or 15	U	N	N	E	D	D	C
2+2U or	16	U	N	N	E	D	D	C
3+2U or	6+U	U	U	U	N	E	D	C
4+2U or	7+U	U	U	U	N	E	D	D
5+2U or	8+U	U	U	U	U	N	E	D
	9+U	U	U	U	U	N	E	E
	10+U	U	U	U	U	U	N	E
	11+U	U	U	U	U	U	N	N
	12+U	U	U	U	U	U	U	N
All other results		U	U	U	U	U	U	U

7.4 DO THE AGGREGATION RULES DELIVER THE INTENDED WEIGHTS OF THE CORE EXAMINATION AND MODULES?

One issue which immediately arises for any examination which uses grade aggregation rules such as those shown in Table 7.1, is that of the relative weights which the examination components are intended to exert within the aggregation process. In the particular case considered here, how do the aggregation rules reflect the intended weightings of 60% and 40% for the core examination and modules respectively? In Chapter 4, it was pointed out that, in general, examination component weights are related to the slopes of the grade boundaries in the score space. In this section, use is made of this idea to develop a technique for addressing the question of intended weighting for any set of grade aggregation rules.

The first step is to re-draw the combination rules of Table 7.1 as a score space. However, before that can usefully be done in the present case, some discussion of the rows corresponding to module sums including ungraded results is required. It can be inferred from these rows that the designers of the aggregation rules took the view that an ungraded result on a module was substantially worse than simply the next point on the scale below Level 4. At best, for results involving 2 ungraded modules, each ungraded module is treated as Level 7 for aggregation purposes; at worst, for the results involving a module sum of 6 (or worse) with one ungraded module, the ungraded module is treated as Level 11. (Strictly, all that can be said is that it is treated as worse than Level 10 since no module sum greater than 16 can be achieved without an ungraded module.) For the 3+U, 4+U and 5+U module sum results, the ungraded module is treated as Levels 8, 9 and 10 respectively. The reasons for these values were not documented by the developers of the aggregation rules but it can be inferred that the intention was, at least partly, to reflect the possibility, given that the modules were completed as coursework, that candidates could avoid ungraded module results by revising and resubmitting their folders and, thus, to motivate candidates to do this. It is clear from Wessex project documents that there was a view that candidates with more than 2 ungraded modules would be unlikely to want to take the core examination and the aggregation rules give an ungraded result for such candidates regardless of their core examination result.

Whatever the detailed rationale for treating ungraded modules so harshly, the effect was clearly to give them greater weight in determining subject grades than would be expected from

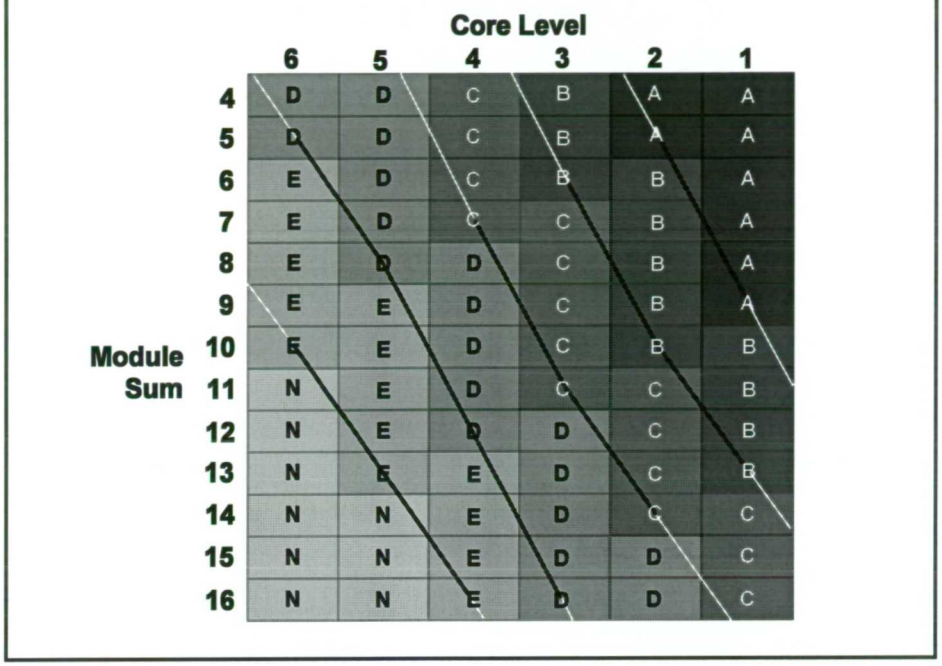
the 40% weighting declared for the modules in the syllabus. Even full marks on the 60% core examination cannot compensate for a very poor performance on more than 2 of the 10% modules. In this sense, then, the first conclusion which can be drawn about the weightings given to the components by the aggregation rules is that they are not the same for candidates with any ungraded modules as for other candidates. The weighting of the modules is increased as the number of ungraded modules increases. In practice, however, of the 147 candidates taking the core examination in Summer 1990 and providing data for this chapter, none had any ungraded modules.

A similar situation applies to candidates who are ungraded on the core examination: they are also ungraded for the examination as a whole. The effect is to give infinite weight to the core examination for candidates who are not graded on it so that even exceptionally good module performances cannot compensate. In practice, of the 147 candidates taking the core examination in Summer 1990 and providing data for this chapter, 29 were ungraded.

The weighting issue which remains thus concerns only the portion of the aggregation rules which relates to candidates achieving a result on all their modules and the core examination and where, therefore, the issue of trade-off between core and modules arises. The score-space analysis outlined in Chapter 4 can now be applied by re-drawing this portion of the aggregation rules as a reduced score space, as in Figure 7.1.

The first important point to note is that the score space has been proportioned as a square. This facilitates the analysis that follows. It corresponds to an assumption that the scales for both the module sums and the core levels span the same range of Chemistry attainment (see Cresswell, 1987a). It is important to note that this assumption is not additional but is implicit in the very notion of the intended weights of the two components being implemented through an aggregation rule matrix. This issue arises here because, unlike the theoretical cases discussed in Chapter 4, the scales used for the two components being combined have different numbers of intervals. This is a common feature of practical examinations.

Figure 7.1
Reduced Score Space for Wessex Chemistry



Lines have been added to the score space shown in Figure 7.1 to represent the average slopes of the subject grade boundaries. The locations of these lines have been defined only by cells which unambiguously define the limits of the particular subject grade; these portions of the lines are shown in black. Thus, the bottom E cell in the Core Level 4 column has been used to anchor a black line because Table 7.1 confirms that this is the lowest E cell in this column. On the other hand, the bottom C cell in the Core Level 1 column has not been used to anchor a black line because Table 7.1 shows that it is **not** the bottom C cell in this column. The black lines have been drawn between the central points of the relevant cells and then extrapolated (in white) to the edges of the reduced score space. Thus, cells in the space which have most of their area immediately below and to the left of a line represent the grade below the grade represented by the cells bisected by the line or having more than half their area immediately above and to the right of the line.

The first significant point to note from the reduced score space diagram in Figure 7.1 is that the central three boundary lines are not straight. The implication of this is that the relative weights of the core and modules are different in different parts of the score space. Two different slopes occur in all the line segments. To what relative weights do these correspond?

The direct way to answer this question would be simply to measure the angles which the lines make with the horizontal axis. Since the space has been drawn as a square, these angles are related to the weights in the way set out in Chapter 4:

$$\tan(\theta) = \frac{W_C}{W_M}$$

where θ is the angle which the boundary line makes with the core level axis, W_C is the weight given to the Core and W_M is the weight given to the Modules.

However, there is also a straightforward analytical approach since,

$$\tan(\theta) = \frac{\frac{L}{N_v} \cdot n_v}{\frac{L}{N_h} \cdot n_h}$$

where L is the length of the sides of the square score space; N_v is the number of levels on the vertical axis, n_v is the number of levels through which the line extends vertically, N_h is the number of levels on the horizontal axis and n_h is the number of levels through which the line extends horizontally,

$$\frac{W_C}{W_M} = \frac{n_M}{n_C} \cdot \frac{N_C}{N_M}$$

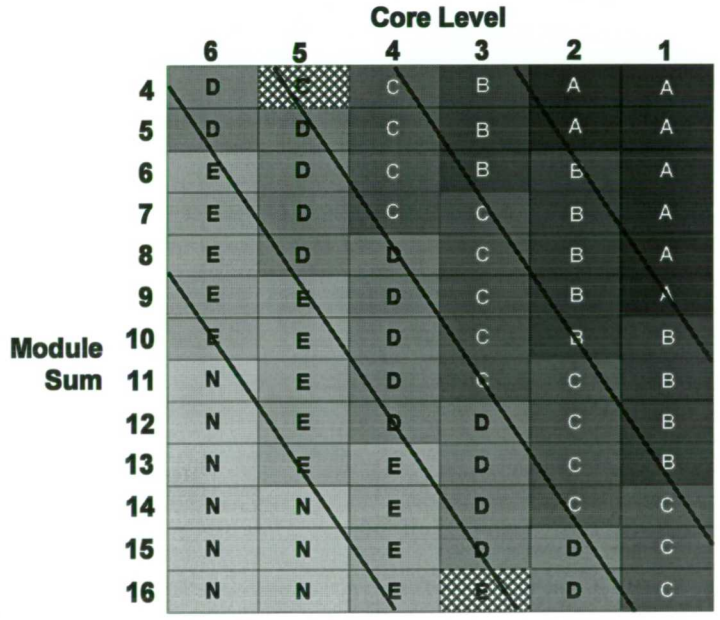
where n_M is the number of Module levels through which the line extends, n_C is the number of Core levels through which the line extends, N_C is the total number of Core levels and N_M is the total number of Module levels. Each line segment shown in Figure 7.1 falls into one of two slopes corresponding to $n_M = 3$ or $n_M = 4$ when $n_C = 1$. Thus, the approximate relative weights given to the core and modules by the aggregation rules are either 58.1 to 41.9 or 64.9 to 35.1.

Because there are 13 module levels and 6 core levels, the intended weights would be perfectly reflected in the aggregation rules if the boundary lines corresponded to $n_M = 13$ when $n_C = 4$. Figure 7.2 shows the score space of Figure 7.1, revised to reflect the intended weights on this basis. At each boundary, the boundary line has been inserted in such a way

as to minimise the number of cells which then correspond to a grade different from that determined by the operational aggregation rules.

The similarity between Figures 7.1 and 7.2 concerning the grade to which each cell corresponds, shows that, in the main, the deviations from the declared intended weights of 60 to 40 for the core and modules respectively are simply an inevitable result of the approximations involved in combining two coarse variables. There are just two cells in the reduced aggregation rules (shown in cross-hatching) which should have corresponded to different grades to reflect the declared intended weights for the core and modules.

Figure 7.2
Revised reduced Score Space for Wessex Chemistry



7.5 THE CANDIDATES' RESULTS

The way in which the aggregation rules operated to produce candidates' results is described in this section. Figures 7.3 and 7.4 show the distribution of core levels and module sums respectively for the 147 candidates whose data were analysed.

Figure 7.3
Core level distribution
(mean = 4.48; SD = 1.89 after coding U=7)

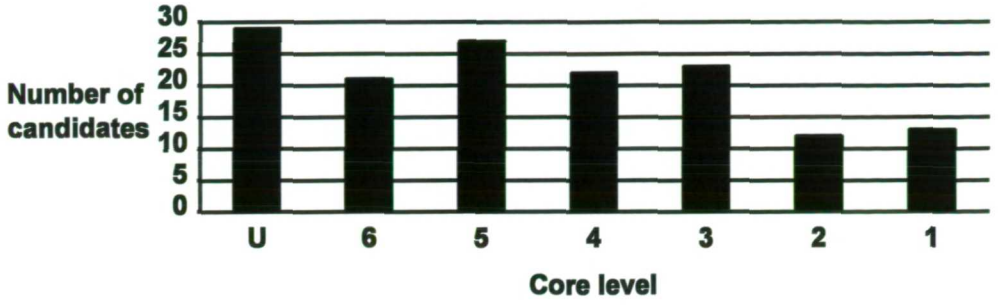
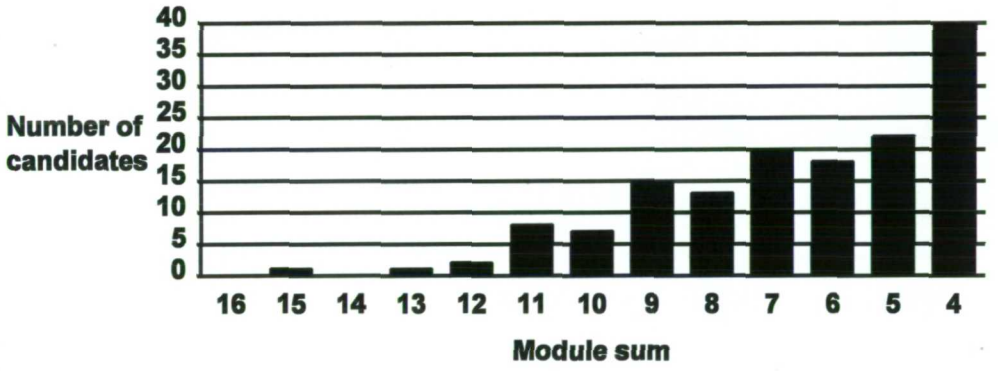


Figure 7.4
Module sum distribution
(mean = 6.58; SD = 2.41)



It can be seen from these figures that the core levels are reasonably conventionally distributed except for the large proportion of *Ungraded* candidates but the module sums are very negatively skewed. Figure 7.5 shows how these data are distributed within the aggregation rule score space (the rows for ungraded modules are not shown since no candidates had any ungraded modules). The coefficient of correlation (coding Core *Ungraded* as Level 7) between the Core levels and module sums is 0.54.

The subject grades were distributed as shown in Figure 7.6. The awarders expressed themselves satisfied, on the basis of their qualitative judgement during the verification meeting (equivalent to conventional awarding Step 4; see Appendix 7.1), that the final grades were, in general, awarded comparably with grades in the Board's conventional A-level Chemistry examination. However, there was some concern expressed during the verification meeting

about the 3 candidates who were ungraded because of their very poor performance on the core examination, despite having module sums of 5 or 6.

Figure 7.5
Distribution of core levels and module sums within the aggregation rule score space

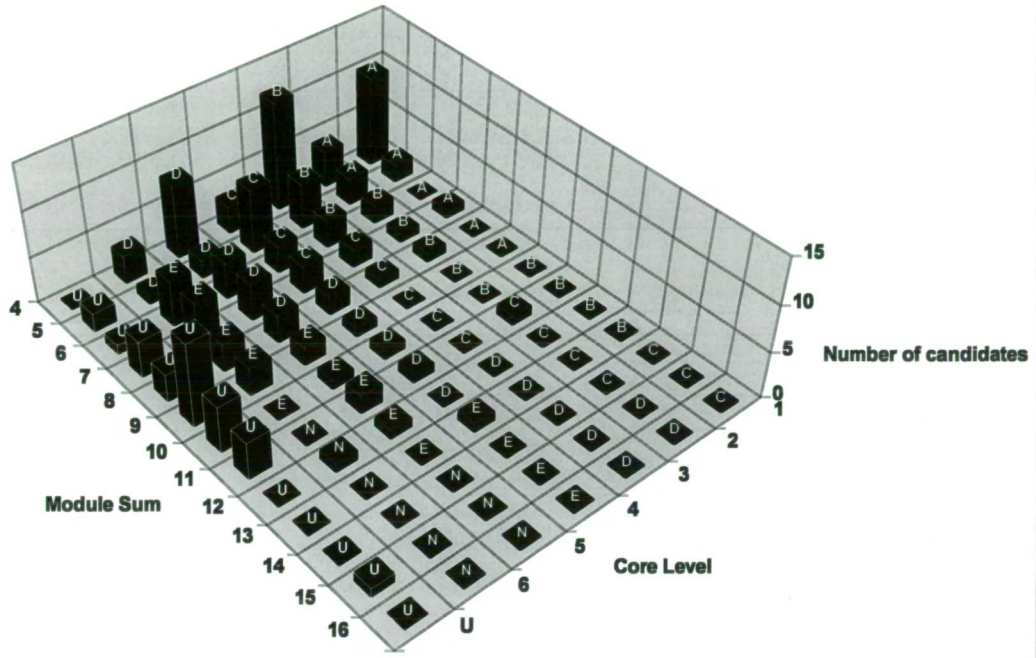
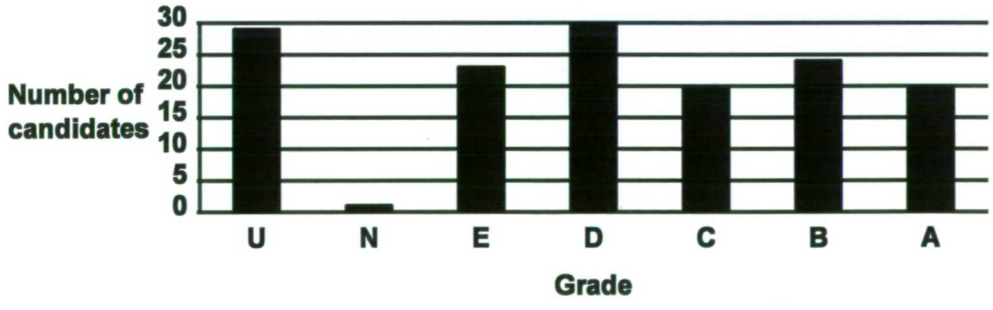


Figure 7.6
Subject grade distribution
(mean = 3.89; SD = 1.96, after coding A=1, B=2,... U=7)



Given the small size and unknown representativeness of the sample of students involved, there are no direct data pertaining to comparability which might challenge the awarders'

collective view that the subject grades were generally appropriate. However, considered in terms of the distributions of grades normally produced by A-level examinations (see, for example, AEB 1995) and the reasonably continuous distributions of levels reported in Figures 7.3 and 7.4, the lack of candidates awarded a subject grade of N, despite many Grade U and Grade E results, is anomalous. This anomaly is caused by the interaction of two effects: the infinite weight of the core examination for candidates who are ungraded upon it and the extent to which the negatively skewed module sums have substantially enhanced most candidates' subject grades compared with their core levels.

It is important to note immediately, however, that the way in which the module results enhance the core levels to produce the subject grades does not imply that the modules exerted an undue weight within the candidates' subject grades. The situation is entirely symmetrical and can equally be viewed as the core examination reducing the module results. It is only because there is a one to one correspondence between the core levels and A-level grades and because an ungraded result on the core examination leads to an ungraded result overall that it seems natural to treat the core levels as some sort of baseline. The enhancing effect of the module results when they are combined with the candidates' core levels arises solely because the candidates apparently performed very much better on the modules than on the core examination. Although it could be that the modules were leniently graded, there is no independent evidence of this and it is axiomatic within any grade aggregation system that the component grades have been correctly awarded. This axiom is required because the aggregation rules, by their very existence, prohibit subsequent adjustment of the standards of work required for the subject grades in the light of knowledge about the accuracy, or otherwise, of the component grades.

As noted in the previous section, it is implicitly assumed in grade aggregation systems that the component grade scales span the same range of attainment. The leading diagonal of the score space thus identifies equivalences between the components' grade scales. Since the core levels were anchored to the normal A-level grade scale (see Section 7.3) it can be deduced, given the implicit assumption of equal attainment spans, that the module sums were related to grades in the way shown in Table 7.2. Thus, over 87% of the candidates were

awarded the equivalent of a Grade C or better on their modules (see Figure 7.4). In the light of this, the subject grade distribution does not seem unduly negatively skewed.

Table 7.2
The relationship between module sums and A-level grades implied by the aggregation rules and intended weights

Module sum	Grade
4	A
6	B
9	C
11	D
13	E
16	N

7.6 THE ACHIEVED WEIGHTS OF THE CORE EXAMINATION AND MODULES

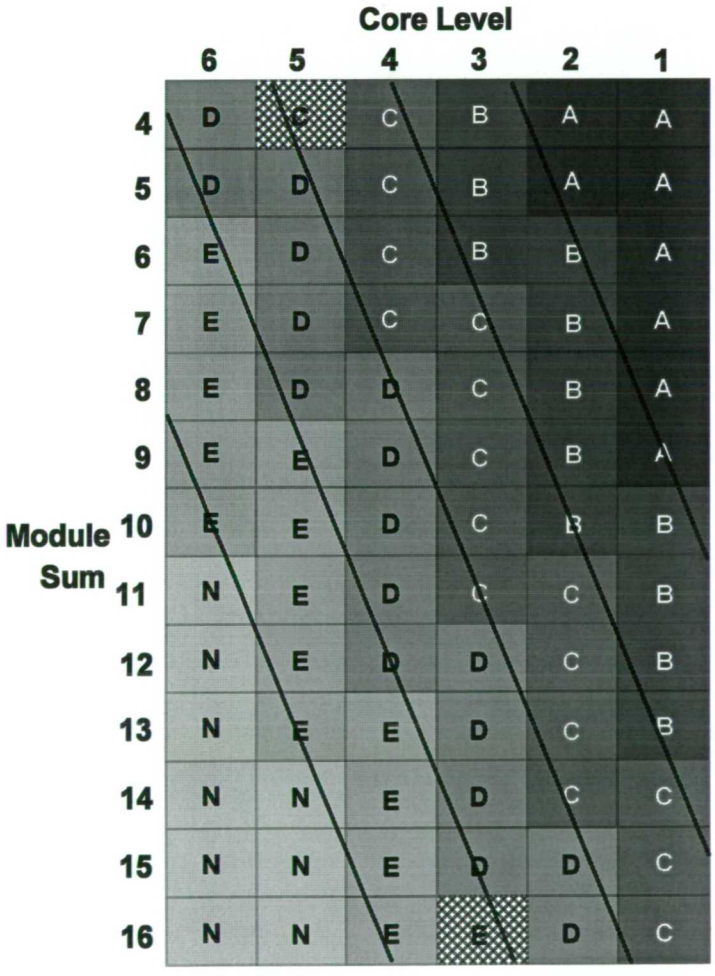
Even though, as shown above, the considerable enhancing effect of the module results upon the core results does not imply any deviation from the intended weights discussed in Section 7.4, the question remains of how much each component contributes, in practice, to the final rank ordering of candidates. In more conventional mark aggregation systems, this question is answered by computing the *achieved weights* of the components.

One way of defining the achieved weight of components within a grade aggregation system, which has the benefit of using the same theoretical ideas as the analysis of intended weight carried out in Section 7.4, is to apply the conception of achieved weight proposed in Cresswell (1987a). This defines the relative weights of the components in terms of the rate of exchange of component marks or, in this case, component grades, within the aggregate. The easiest way to apply this definition of achieved weights in the present context, is simply to rescale the reduced score space to be square in terms of the standard deviations of the component grades, rather than their ranges (see Chapter 4). The slopes of the grade boundary lines in the rescaled score space then reflect the achieved weights of the components as defined by Cresswell (1987a). This has been done in Figure 7.7, using the standard deviations of the

levels for the 118 candidates who lie within the reduced score space: 1.59 for the core levels and 2.15 for the module sums. The analysis developed in Section 7.4 can then be applied.

As might be expected from the ceiling effect evident in the distribution of module sums (Figure 7.4), the contributions of the components to the final rank ordering of candidates do not reflect the intended weights of 60% for the core examination and 40% for the modules. The relative weights implied by the subject grade boundary lines in Figure 7.7 are 70.6 to 29.4.

Figure 7.7
Score space proportioned for the analysis of achieved weights



There are several other definitions of achieved weight in the literature and the most commonly used one is the component-with-subject total covariance (Adams and Murphy, 1982). Rather than being concerned with trade-off rates between the components, this definition partitions

the subject variance into portions contributed by each of the components. This definition is inappropriate in the case of grade aggregation because the subject grades are not simply an arithmetic combination of the component grades and the subject variance cannot therefore be accounted for entirely in terms of the component-with-subject total covariances.

However, the conventional component-with-subject total covariance measure of achieved weight consists of two terms: one for the unique variance of the component scores and one for half the covariance of the component with the other components. A similar approach can be applied to grade aggregation systems by means of multiple regression analysis relating the subject grades, as the dependent variable, to the component grades, as independent variables.

If such an analysis is done, the proportion of the variance in the subject grades which can be associated uniquely with Component i is given by:

$$U_i = R^2 - \left(R_{\tilde{i}}\right)^2$$

where R is the multiple correlation coefficient when all the components are included in the analysis and $R_{\tilde{i}}$ is the multiple correlation coefficient when Component i is omitted from the analysis. The proportion of the variance in the subject grades which can be associated with the component covariances is given by:

$$C = R^2 - \sum_{i}^{n} U_i$$

where n is the number of components, Then, the relative percentage weights of the components can be defined as the solutions of:

$$\frac{W_i}{U_i + C/n} = \frac{W_j}{U_j + C/n} = \ \ldots \ = \frac{W_n}{U_n + C/n} \quad \text{and} \quad \sum_{i}^{n} W_i = 100$$

This definition is conceptually identical to the conventional component-with-subject total covariance definition in the case of two components but differs from it for more than two components because, in such cases, under the conventional definition the common variance is not, necessarily, partitioned equally between all the components.

This definition of the achieved weight of components in grade aggregation systems has been applied twice in the present case: once for the complete data set including the candidates who were ungraded on the core examination and once excluding those candidates. The results are shown in Table 7.3.

Table 7.3
Results of multiple regression weighting analysis

Including candidates ungraded on the Core		
R^2 = 0.95	Percentage weights	
$(R_2)^2$ = 0.92	Core examination	Module sum
$(R_1)^2$ = 0.45	74.3	25.7
Excluding candidates ungraded on the Core		
R^2 = 0.96	Percentage weights	
$(R_2)^2$ = 0.88	Core examination	Module sum
$(R_1)^2$ = 0.36	77.1	22.9

It is clear that, by this measure, the effect of the module results upon the rank-ordering of candidates within the aggregate is again shown to be considerably less than the intended weight of 40%. This reflects the inevitable bunching of the heavily skewed module sum distribution (see Figure 7.4).

In one sense, since it is axiomatic under grade aggregation that the core and module levels were accurately awarded, the discrepancies reported here between achieved and intended weights are simply a reflection of the relative similarity of the candidates' performances on the modules, compared with their more disparate performances on the core examination. All analyses of achieved weights which use candidate data share the characteristic that they reflect the interaction between the assessement instruments and candidates, rather than characteristics of the assessment instruments *per se*. It is, therefore, possible that a different group of candidates might perform in such a way as to give different measures of achieved weights.

The consequences of the candidate-dependent nature of most measures of achieved weight are discussed at some length in Cresswell (1987a). Here it will simply be noted that, in the present case, it would not **necessarily** be appropriate to revise the assessment criteria used to assess the modules in order to reduce the ceiling effect bunching in the distribution of module sums and thus increase the achieved weight of the modules in the aggregate. Like the measures of achieved weight used in conventional aggregation, the measures of achieved weight for grade aggregation systems which have been developed in this section should not be interpreted prescriptively but, rather, as descriptions of the way in which the examination functioned for a particular group of candidates.

7.7 A JUDGEMENTAL COMPARISON WITH CONVENTIONAL GRADING PROCEDURES

7.7.1 The results re-graded conventionally

As observed above, the distribution of subject grades produced by the grade aggregation rules appears anomalous with respect to the proportion of candidates awarded Grade N. In addition, as discussed in Chapter 4, there are inevitable approximations involved in grade aggregation compared with conventional awarding on a total subject mark scale. To explore these two effects, the marks from the core assessments and modules were added together (using suitable scaling factors to achieve the specified intended weights for all the components) and the resulting total mark scale was then partitioned by conventional grade boundaries. The grade boundaries were positioned so as to give, as far as possible, proportions of candidates at each of the key grade boundaries (A/B, B/C and E/N - see Chapter 5) which were the same as those produced by the grade aggregation rules. The remaining conventional grade boundaries were then positioned using the normal interpolation rules for A-level examinations (see Appendix 5.1). The results of this analysis are presented in Table 7.4.

Table 7.4
Comparison of subject grades from aggregation rules with grades awarded on aggregated marks

Grades from aggregation rules	Grades from mark aggregation							
	A	B	C	D	E	N	U	Total
A	18	2	-	-	-	-	-	20
B	2	18	3	1	-	-	-	24
C	-	5	13	2	-	-	-	20
D	-	-	10	17	3	-	-	30
E	-	-	-	4	15	3	1	23
N	-	-	-	-	-	-	1	1
U	-	-	-	-	4	8	17	29
Total	20	25	26	24	22	11	19	147

From Table 7.4, it can be seen that there are 49 candidates (exactly one third) who receive different grades, depending upon the awarding method used. Given the consequences for candidates of their public examination results (Chapter 2), this is clearly a large proportion of grade changes (although it is comparable with the proportions of grade changes reported by Good and Cresswell (1988a) which can occur as a result of the differences of judgement between different teams of awarders using conventional awarding procedures; see Chapter 3). It therefore seems appropriate to ask if one method or the other produces results which were more in keeping with the views of subject specialists about the quality of the candidates' work. A small study was carried out to explore this question.

7.7.2 The collection of the judgements

The entire work of 21 of the candidates whose grades using grade aggregation were different from their grades using mark aggregation was evaluated by 8 suitably qualified judges. The candidates involved are indicated by the light shading in Table 7.4. The selected candidates were at the key Grade A, Grade B and Grade E boundaries because it is these grades which are set at awarding meetings (see Appendix 5.1) and for which examiners therefore have

experience of evaluating scripts. The judges comprised 3 Wessex Project representatives, 2 members of the board's Standing Advisory Committee for Chemistry and 3 experienced A-level Chemistry examiners who were not involved in examining Wessex Chemistry.

The judges compared the work of each selected candidate with that of reference candidates who were awarded the same grade by both methods. For example, at the Grade A boundary there were two candidates who were awarded a Grade A by the aggregation rules but a Grade B by mark aggregation and two more who were awarded a Grade B by the aggregation rules but a Grade A by mark aggregation. The work of these four candidates was compared with that of two candidates who each just merited a Grade A from both aggregation methods. The awarders were asked to indicate if, in their judgement, each of the four candidates deserved a definite Grade A, a bare Grade A, not quite a Grade A or much less than Grade A. Similar judgements were made at the other two key boundaries.

To help the judges to form a reasoned evaluation of each candidate's work, they worked in pairs. There were four pairs. Initially, three pairs judged the work of 5 candidates and one pair judged the work of 6 candidates: 1 at the A/B boundary, 2 (3 for one pair) at the B/C boundary and 2 at the E/N boundary. At each boundary, the particular candidates whose work was judged by each pair were chosen randomly without replacement from each relevant cell in Table 7.4 and the judges were not told the grades awarded to the candidates by either method. Once the initial sets of work had been judged, each pair of judges passed them on to the pair on the next table. In this way, two pairs judged the work of each selected candidate. Unfortunately, there were 3 candidates for whom all the module work was not available. In these cases, the judges were asked to make the best judgement they could. In the event, only for one of the affected candidates did the judges express concern about the difficulty of reaching a judgement in which they had confidence.

The judges were provided with the core question papers and practical marks of the selected candidates, the module assessment criteria, documents specifying the procedures they should follow and forms upon which to record their decisions (copies of the study documents form Appendix 7.2). The module work of all the candidates had been annotated before the

meeting by the moderator for Wessex Chemistry so that the parts relevant to each assessment criterion were identified.

At each of the three grade boundaries, the judges were given the following instructions:

Step 1

Begin by studying the work of the reference candidates for the grade which you are considering. Discuss it with your partner. The intention is that you should assimilate the standard which was applied in the Summer and which is represented by the reference candidates. This will take a considerable time to do properly. Note that you are not being asked to re-grade the reference candidates so try to put aside any reservations you may have about the grades awarded to them.

Step 2

When you are satisfied that you appreciate the standard represented by the reference candidates, study the work of the first selected candidate. Compare it with that of the reference candidates. Work with your partner, discussing any points of interest and/or significance in the selected work, but form your own judgement of the grade which it should earn. Record your individual judgement on Form I (pink) for the grade in question.

Step 3

When you and your partner have each formed an individual judgement about a particular selected candidate's work, see if you agree. If you do agree, enter your collective judgement on Form C (yellow) for the grade in question. If you do not agree, discuss the candidate's work again. There may be some feature of it (or of the reference work) which one of you has missed or under-valued. If you reach a collective judgement as a result of this discussion, record it on Form C (yellow). Do not feel obliged to reach agreement, however, and if you cannot agree about a particular candidate's work, record your disagreement on Form C (yellow).

Step 4

Consider the next selected candidate in the same way.

The meeting proceeded smoothly and productively. The judges seemed to fully understand and recognise the value of their task; they approached it in a co-operative and professional way. The meeting began at 10:30 and ended at about 3:30 pm. In that time, each pair of judges scrutinised the work of 10 or 11 selected candidates. At the end of the meeting, a brief plenary session was held in which the judges were asked to give their impressions of the day's work.

Among the comments made by the judges during the final plenary session, the following two points seem particularly worth noting. First, the judges said that they found it very much easier to judge the work of the best candidates than that of those obtaining Grade E or less and, second, the judges felt that the practice of working in pairs, with each pair considering the work of only a small number of candidates, enabled more considered judgements to be made than were possible in normal awarding meetings.

7.7.3 The evaluative judgements

The results of the study are summarised in Table 7.5. In no case were the judges within a pair unable to agree about the grade which a candidate's work merited so the data have been analysed as 42 independent judgements. When the grade awarded in the present study was not the higher of the two other grades, it was assumed to be the lower if the judges indicated that it was *not quite* worthy of the higher. If the judges indicated that the work was *clearly* not worthy of the higher grade, it was taken to merit the grade two grades below the higher.

7.7.4 Analysis and interpretation

The first question to ask of the data in Table 7.5 is whether the agreement between the pairs of judges is sufficient to conclude that they were making meaningful distinctions between the candidates' work. As described above, each pair of judges could put each candidate's work in one of four categories: definitely Grade x, barely Grade x, not quite a Grade x or much less than Grade x. Figure 7.8 shows, for a pair of judges considering the same candidate's work, which of the possible outcomes indicate agreement about the candidate's grade.

Thus, of the 16 possible outcomes, 6 produce agreement about the grade which the candidate's work ahould be awarded. This can be used as the basis of a binomial test (Siegel and Castellan, 1988) of the significance of the agreement observed between the pairs of judges. For the data as a whole, the judges agreed on 14 out of 21 occasions which represents significant agreement (probability level < 0.01) compared with what would be expected to arise by chance if the judges were not making meaningful classifications of the work.

Table 7.5
The judgements of the selected candidates' work

Candidate	Aggregation rules grade (AR)	Mark aggregation grade (MA)	Judgements				No. agreeing with AR	No. agreeing with MA
			Pair 1	Pair 2	Pair 3	Pair 4		
1	A	B	A	A			2	0
2	A	B		A	B		1	1
3	B	A			B	B	2	0
4	B	A	B			B	2	0
5	B	C	C	C			0	2
6	B	D		C	B		1	1[2]
7	B	C			B	B	2	0
8	B	C	C			B	1	1
9	C	B	B	B			0	2
10	C	B		C	C		2	0
11	C	B			C	C	2	0
12	C	B	B			B	0	2
13	C	B	C	B			1	1
14	E	N	E[1]	N[1]			1	1
15	E	N		N	N		0	2
16	E	N			U	N	0	2[2]
17	E	U	E			E	2	0
18	U	E	E	E			0	2
19	U	E		E	E		0	2
20	U	E			N	U	2[2]	0
21	U	E	N			N	2[2]	0

1 The judges expressed reservations about this judgement because all the candidate's modules were not available.

2 For these cases, the judges' grades have been treated as agreeing with the lower of the two original grades because they were asked to say whether or not candidates' work merited the higher grades.

From Table 7.5 it can be seen that the judges confirmed the aggregation rule grades on 23 occasions and the mark aggregation grades on 19 occasions. This 23:19 split is not significantly different from an even (21:21) split ($p < 0.65$ using the continuity-corrected two-tailed binomial test for samples of more than 25 as given by Siegel and Castellan, 1988). If the annotated cases in Table 7.5 are excluded from the analysis, the split is 17:15 which is also not significantly different from an even split ($p < 0.85$ using the same test). The

conclusion is therefore that, in general, the judges had no significant preference for one method of aggregation over the other.

Figure 7.8
The outcomes (shaded) which produce agreement in Table 7.5 from two pairs of judges considering the same candidate's work

	Definite Grade x	Barely Grade x	Just < Grade x	Much < Grade x
Definite Grade x	▒	▒		
Barely Grade x	▒	▒		
Just < Grade x			▒	
Much < Grade x				▒

On the matter of the Core examination hurdle, of the eight evaluative judgements of work which was ungraded by the aggregation rules, only one produced a U. This difference is statistically significant ($p < 0.01$ using the one-tailed Kolmogorov-Smirnoff test as given by Siegel and Castellan, 1988) and indicates that the judges did **not** agree with the consequences of the hurdle which gave candidates who were ungraded on the Core examination an ungraded result for the subject as a whole.

7.8 IN CONCLUSION

This chapter has reported a detailed case study of one particular use of grade aggregation. During the study, new approaches to the analysis of intended and achieved weights were developed which, for the first time, can be applied to grade aggregation systems in general.

The study also included a comparative investigation, based upon holistic qualitative evaluations of candidates' work, of the grades awarded by grade and mark aggregation methods. It appears that, in the present case at least, the theoretical arguments against grade aggregation which are set out in Chapter 4 were insufficient to create systematic practical

effects large enough to be of concern to suitably qualified judges. However, the use of a hurdle to give infinite weight to one of the components being aggregated gave results which differed significantly from the judges' holistic evaluations of the affected candidates' work. Interestingly, the possibility of using similar hurdles is sometimes quoted as an advantage of strong criterion-referencing (see, for example, Wiliam, 1995b) and the next chapter reports a case study of such an approach. The conclusion from the small study reported in this chapter is that there is clearly a need for further empirical work, in a variety of different examinations, before a judgement can be made as to whether, in practice, the theoretical defects of grade aggregation discussed in Chapter 4 outweigh its benefits in terms of transparency for candidates and teachers.

CHAPTER 8

AN EXPERIMENT IN STRONG CRITERION-REFERENCING

"'The thing can be done,' said the Butcher, 'I think',
The thing must be done, I am sure,
The thing shall be done! Bring me paper and ink,
The best there is time to procure.'"
The Hunting of the Snark - Lewis Carroll

8.1 INTRODUCTION - THE BACKGROUND TO THE EXPERIMENT

This chapter reports an experiment using awarding procedures which are *strongly criterion referenced* in terms of the analysis in Chapter 3. The experiment was conducted in late 1992 and early 1993 as part of pilot work for new GCSE examinations in Mathematics which were to be introduced in Summer 1994[1].

New GCSE examinations were required in Summer 1994 as part of the introduction of the National Curriculum in England and Wales. 1994 was the first year in which pupils in the final year of compulsory education had followed the National Curriculum throughout the preceding two years (in National Curriculum parlance these two years are known as Key Stage [KS] 4). At the time of the pilot work, the government agency then responsible for regulating GCSE Examinations (The *School Examinations and Assessment Council, SEAC*) had imposed a number of novel requirements on GCSE examinations at the end of KS4 concerning the data which were to be reported for each candidate. In particular, for each candidate the examinations were required to produce a profile of results across the so-called *Attainment Targets* which partitioned each subject into narrower domains, as well as an overall result for the subject as a whole. Both the Attainment Target (AT) results and the Whole Subject result were to be reported in terms of ten levels for which, in each subject, descriptors had been written by a National committee of subject specialists to define supposedly progressively greater degrees of attainment.

---

[1] The work reported in this chapter was part of a programme of GCSE pilot work in Mathematics, Science and English, done under the auspices of the *Joint Council for the GCSE*. The permission of the Joint Council to report it here is gratefully acknowledged. The Mathematics pilot study was directed by the author and carried out in collaboration with Dr Hope Macdonald whose major contribution to the study I should also like to acknowledge. The findings which are reported in this chapter were first reported in Cresswell *et al* (1993); other important findings from the study are given in Macdonald (1993).

The School Examinations and Assessment Council was also responsible for statutory assessment of 14 year olds at the end of Key Stage 3 (KS3) and was attempting to use strongly criterion-referenced procedures for that purpose. Specifically, novel aggregation procedures were being piloted in which, broadly speaking, a student had separately to exceed a mastery score on assessment material identified in National Curriculum documentation with each National Curriculum level, before that level could be awarded. The rationale for this approach was the usual strong criterion-referencing one: the hope that valid inferences concerning the details of students' attainments would be possible from a simple summary score (see Chapter 3). In 1992/3, the use of similar procedures, for the same strongly criterion-referenced purpose, was pressingly advocated by SEAC for the new GCSE examinations at KS4.

In fact, partly as a result of the experimental work reported in this chapter, by 1994 the use of strongly criterion-referenced procedures for National Curriculum assessments had been *all but abandoned*. (Regrettably, few of the data - for example, Ruddock *et al*, 1993 - on the failure of strongly criterion-referenced procedures in pilot National Curriculum assessments have been published.) Following the review of the National Curriculum (Dearing, 1994) GCSE examinations never reported AT profiles and continued to report candidates' Whole Subject attainments on the original scale of GCSE grades, rather than levels, using conventional awarding procedures to do so. More conventional standard setting procedures were also adopted for Key Stage 3. However, as noted in Chapters 1 and 3, successive attempts have been made in Britain to achieve strong criterion-referencing since the early 1980s and, at the time of writing, this failed, but seductive, idea is apparently still alive (Hutchison and Schagen, 1994; Massey, 1995).

Against this background, since one of the purposes of the present work (see Chapter 1) is to explore possible alternative approaches to awarding public examination grades, it is necessary to look at strong criterion-referencing in some detail, despite its recent eclipse for the purposes of making National Curriculum assessments. Major theoretical problems afflicting strong criterion referencing were set out in Chapters 3 and 4. The experiment reported in this chapter tried it out in practice, in the context of public examinations, using

procedures based upon those being used in 1992/3 for National Curriculum assessments at Key Stage 3. The strongly criterion-referenced procedures were compared with conventional awarding procedures which were fundamentally those which are the main focus of Chapters 5 and 6, but which had been adapted to produce a profile of Attainment Target levels as well as an overall level. The experiment reported here is the only known direct comparison between strongly criterion-referenced procedures and conventional examination awarding procedures.

8.2 OUTLINE OF THE EXPERIMENT

The experiment involved a sample of candidates sitting Mathematics examinations which were structured so as to allow strongly criterion referenced awarding procedures like those being used at Key Stage 3 to be used alongside more conventional awarding procedures. The results of the two procedures were then compared with each other and with holistic estimates of Mathematics attainment provided by the sample candidates' teachers.

The experiment was conducted over a six-month period. The papers and marking schemes were prepared during September and October 1992. Like all GCSE Mathematics examinations, the experimental papers were organised in three tiers which differed in difficulty. The candidates sat the papers, under normal examination conditions, during the last two weeks of November and the first week of December. Within each centre, the written papers were administered on separate days, with no more than seven days between the two dates; however, the exact dates of the examination differed across centres. The papers were marked during the ensuing four weeks, the mark data were processed and an awarding meeting was held during the first week of February 1993.

8.3 THE SAMPLE

All of the sample candidates were Year 11 GCSE students following the Southern Examining Group's *Mathematics B* syllabus. Centres were approached in either July or September 1992 about participating in the pilot examinations. The final sample consisted of 489 candidates

from 5 centres. One centre was a mixed independent school and the other 4 centres were LEA-maintained mixed comprehensive schools.

The extent to which the sample represented all candidates and centres who enter for GCSE Mathematics examinations was not of prime importance to this study. The main requirement was for candidates spanning the full range of GCSE attainment so that the awarding procedures could be applied at all grades. The independent school in the sample had prepared and entered its 32 candidates for the November 1992 SEG Mathematics B examination, all the other candidates were due to be entered for the following summer's examination. However, there was no evidence from the candidates' scripts that they had not covered all the material assessed and nor were there any comments from their teachers to this effect.

Prior to entering individual candidates for the pilot, the teachers were given information about the three tiers of assessment and guidance concerning the officially expected links between GCSE grades and National Curriculum levels (TGAT, 1988). The teachers were asked to consider each candidate's likely GCSE grade, *at the time of the experiment*, to translate this grade into a National Curriculum level and then to enter candidates expected to attain Levels 4 and 5 for Tier 4-6, those expected to attain Levels 6 and 7 for Tier 5-8, and those expected to attain Levels 8, 9 or 10 for Tier 7-10. The size of the sample for each tier was as shown in Table 8.1.

Table 8.1
Sample sizes for the three tiers

Tier	Number of candidates
4 - 6	86
5 - 8	275
7 - 10	128
Total	489

8.4 THE EXPERIMENTAL EXAMINATIONS

The experimental examinations were prepared by adapting and revising the GCSE KS4 specimen papers produced by the Southern Examining Group (SEG) to accompany the new 1994 syllabus. Full details of the rationale behind the design of the experimental examinations, and the related issues which emerged during the main pilot study, are reported by Macdonald (1993). An extensive account of these matters is not necessary for the purposes of this chapter but the main structural features of the experimental examinations are described in this section.

8.4.1 The Tiers and Level Weightings within them

As noted in Section 8.1, at the time of the pilot work the 1994 GCSE examinations were intended to report candidates' attainment in terms of National Curriculum levels, rather than grades. Like all GCSE Mathematics examinations, the experimental examinations were organised into three tiers of papers, these tiers differing in difficulty and giving access to different ranges of levels. Since each National Curriculum level was defined by a number of Statements of Attainment (SoAs), and each question in the experimental examination was intended to address one or more specific SoAs, the marks allocated to each level in the experimental papers defined the tiers of assessment. For each tier, the structures of marks per level were as shown in Table 8.2.

Table 8.2
The experimental examination: Marks per Level

Tier	Marks per level							Total Marks
4 - 6	L4	L5	L6					
	37	**44**	**39**					**120**
5 - 8		L5	L6	L7	L8			
		30	**60**	**60**	**30**			**180**
7 - 10				L7	L8	L9	L10	
				32	**55**	**62**	**31**	**180**

Each tier comprised two written papers and a coursework assessment with intended weights of 40%, 40% and 20%, respectively.

8.4.2 The written papers

Each of the written papers covered just three levels and as a result, in Tier 5-8 and in Tier 7-10 where there were four targeted levels, the two written papers were stepped such that the lowest and highest levels of the tier were not included in the same paper (see Table 8.3). The papers in Tier 4 - 6 were each intended to take candidates 1 hour to complete, those in Tier 5 - 8, 1.5 hours and those in Tier 7 - 10, 2 hours.

Table 8.3
Levels Targeted in Papers

Tier	Paper	Targeted levels
4 - 6	1	4 5 6
	2	4 5 6
5 - 8	1	5 6 7
	2	6 7 8
7 - 10	1	7 8 9
	2	8 9 10

The experimental papers represented a considerable departure in the structuring of GCSE papers since they were divided into sections, each of which contained questions intended to assess a particular Attainment Target (AT). The questions within each AT section were ordered according to the National Curriculum level addressed, beginning with those aimed at the lowest levels, rather than according to question difficulty as judged by the examiners. This sectionalised approach was adopted to facilitate the extraction of a score for each AT which, at the time, was one of the major requirements of the proposed National Curriculum GCSE examinations.

The Attainment Targets for National Curriculum Mathematics were:

Ma1 Using and applying mathematics
Ma2 Number
Ma3 Algebra
Ma4 Shape and space
Ma5 Handling data

The intended weight of each AT was 20%, as specified in the Mathematics National Curriculum. The experimental written papers assessed Ma2, Ma3, Ma4 and Ma5; Ma1 scores were represented in the study by teachers' estimated coursework marks.

8.4.3 Coursework (Ma1) marks

Each pilot candidate's teacher was asked to provide, on the basis of SEG *Mathematics B* coursework completed by December 1992, an unmoderated coursework mark which could be used to represent the candidate's Ma1 mark in the KS4 pilot examination. The SEG *Mathematics B* syllabus had a centre-assessed component which was weighted at 20% and marked on a scale where each mark related to one of the GCSE grades.

Table 8.4
Coursework Level Boundaries

GCSE GRADE	COURSEWORK MARK	NC LEVEL
	20	10
A	19	9
B	18	8
	17	
B	16	8
C	15	7
	14	
C	13	
D	12	7
	11	6
D	10	
E	9	
	8	
E	7	6
F	6	5
	5	
F	4	5
G	3	4
	2	
G	1	4
U	0	0

Using the officially expected relationship between GCSE grades and National Curriculum levels (TGAT, 1988) the level boundaries shown in Table 8.4 were fixed prior to the awarding meetings. The coursework itself was not requested from centres to lessen the demands made upon them by participation in the experiment.

8.5 TEACHERS' ESTIMATES

The candidates' teachers were also asked to provide an estimate of the GCSE grade each candidate would achieve, if that candidate were to sit standard GCSE examination papers **on the dates scheduled for the experimental papers**. It was emphasised that this grade should not be a prediction about performance in the Summer 1993 examination. The estimated grades were made in terms of half-grades, that is, using a + sign to refer to the top half of the range for each grade and a - sign to designate the bottom half of the grade range. They were obtained for all candidates and transformed to NC levels, on the basis of the officially expected relationship between grades and levels, as shown in Table 8.5.

Table 8.5
Relationship between Forecast Grades and NC levels

Forecast GCSE grade	National Curriculum Level
A+	10
A–	9
B+	8
B–	8
C+	7
C–	7
D+	7
D–	6
E+	6
E–	6
F+	5
F–	5
G+	4
G–	4

8.6 CONVENTIONAL AWARDING

In outline, the procedure used to award the experimental examination conventionally was as follows:

1. For each Attainment Target on the written papers, each candidate was given a total mark which was the sum of the marks they obtained on all the questions addressing that AT.

2. Within each AT on the written papers, boundary marks were set conventionally (see Chapter 5) to form successive ranges of marks corresponding to the AT levels.

3. For each Attainment Target, each candidate was awarded the level which was associated with the range of AT marks within which his or her AT mark lay.

4. Each candidate's AT marks, including the teacher's coursework marks, were then summed, weighted as necessary to give the ATs equal weight, to form a total mark for the subject as a whole.

5. The AT boundary marks were used, including those for the coursework marks (see Table 8.4), to determine a series of boundary marks on the total mark scale which defined successive ranges of marks corresponding to the Whole Subject levels. The method used to produce the Whole Subject boundary marks was the conventional one of taking the lower of the two subject marks produced by the *addition* and *percentile* grade boundary combination methods (see Chapter 4).

6. Each candidate was then awarded the Whole Subject level which was associated with the range of Whole Subject marks within which his or her total mark lay.

The various judgemental decisions required by this process were made at an awarding meeting, constituted conventionally to include the examiners who set and marked the papers and some additional subject specialists. Full details of the proceedings of the awarding meeting are given by Macdonald (1993) who reported that, in the main, "the awarding meeting progressed smoothly" and the examiners were satisfied with the decisions which they had made.

The summary statistics and conventional level boundaries for the Attainment Target marks are given in Appendix 8.1. The results of the conventional awarding process were as shown in Table 8.6.

Table 8.6
Distributions of conventionally awarded levels
(percentage of candidates in each awarded level)

Tier (no. of cands)	Level	Ma1[1]	Ma2	Ma3	Ma4	Ma5	Whole Subject
4 - 6 (86)	U	1.2	27.9	22.1	48.8	39.5	22.1
	4	32.6	57.3	58.1	37.2	41.9	50.0
	5	36.0	12.8	18.6	12.8	17.4	26.7
	6	29.1	2.3	1.2	1.2	1.2	1.2
5 - 8 (275)	<5 = U	5.9	0.4	1.1	11.3	7.3	1.5
	5	12.4	30.9	42.5	36.7	40.7	29.8
	6	48.0	46.5	43.3	31.6	36.7	47.6
	7	29.1	15.3	10.5	17.1	13.1	17.1
	8	4.7	6.9	2.5	3.3	2.2	4.0
7 - 10 (128)	<7 = U	16.4	0.0	0.0	5.5	11.7	1.6
	7	35.2	25.0	46.1	26.6	30.5	37.5
	8	36.7	62.5	32.8	31.3	35.2	39.1
	9	7.0	12.5	15.6	30.5	22.7	18.8
	10	4.7	0.0	5.5	6.3	0.0	3.1

[1] Ma1 levels awarded as described in Section 8.4.3

8.7 THE STRONGLY CRITERION-REFERENCED PROCEDURE

The strongly criterion-referenced procedure used to award the experimental examination was based, as closely as possible, on the one being used nationally, at the time of the experiment, for pilot statutory National Curriculum assessments at Key Stage 3. In outline, it was as follows:

1. For each Attainment Target, marks were totalled for each candidate's work **within** each level (but were not added up across levels).

2. Within each AT **level** separately, the candidate's marks were compared with a pre-determined mastery score of 66.6%. (This value is quite low from a strongly criterion-referenced perspective because it means that the only inference that can be drawn from a candidate's reported level is that they are competent in an unspecified two-thirds of the assessment domain represented by material at that level.)

3. For the AT as a whole, the candidate was awarded the highest level at which his or her marks exceeded the mastery score, regardless of whether he or she had exceeded the mastery scores at lower levels. (See Chapter 4 for a full discussion of the rationale for, and issues arising from, aggregation methods of this type.)

4. Candidates' Whole Subject levels were determined by forming a weighted average of their AT levels, rounded to the nearest integer. (Certain arbitrary rules were required to determine the Whole Subject levels of candidates who did not receive a level on one or more ATs. The details and effects of these are discussed in Section 8.9.)

Table 8.7 shows the results of these strongly criterion-referenced procedures for the sample candidates

Table 8.7
Distributions of levels awarded by the strongly criterion-referenced procedure (percentage of candidates in each awarded level)

Tier (no. of cands)	Level	Ma1[1]	Ma2	Ma3	Ma4	Ma5	Whole Subject
4 - 6 (86)	U	1.2	48.8	20.9	91.9	75.6	83.7
	4	32.6	38.4	77.9	5.8	9.3	14.0
	5	36.0	9.3	0	2.3	9.3	2.3
	6	29.1	3.5	1.2	0	5.8	0
5 - 8 (275)	<5 = U	5.9	25.8	58.2	47.3	46.2	56.7
	5	12.4	18.2	6.9	17.1	6.2	9.1
	6	48.0	22.2	18.9	4.0	40.4	14.9
	7	29.1	4.0	13.1	6.2	6.2	17.1
	8	4.7	29.8	2.9	25.5	1.1	2.2
7 - 10 (128)	<7 = U	16.4	28.9	56.3	25.0	43.8	57.0
	7	35.2	17.2	21.1	31.3	18.8	14.8
	8	36.7	22.7	4.7	13.3	18.8	17.2
	9	7.0	3.1	17.2	18.8	18.8	10.9
	10	4.7	28.1	0.8	11.7	0	0

[1] Ma1 levels awarded as described in Section 8.4.3

8.8 THE ATTAINMENT TARGET RESULTS FROM THE TWO PROCEDURES

In this section, the levels produced by the two procedures for Attainment Targets 2 to 5, and reported in Tables 8.6 and 8.7, are compared and evaluated. The next section considers the whole subject levels.

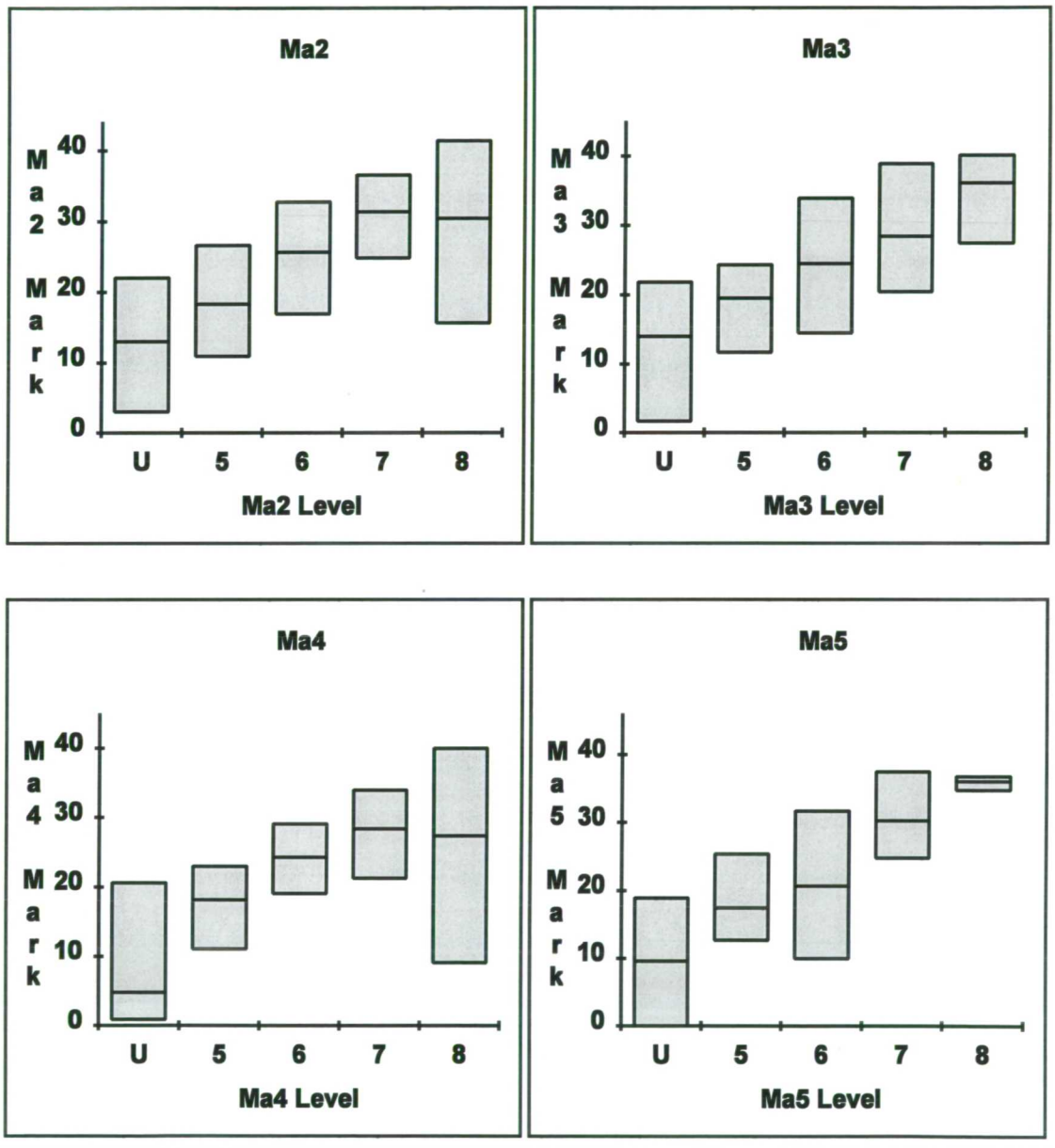
While the distributions of AT levels produced by the conventional awarding procedures are typical of the way in which measured educational attainment is usually distributed, those produced by the strongly criterion-referenced procedure are erratic and have no known theoretical justification. These distributions imply that the attainment of the pilot study's candidates on the ATs is distributed in a way which is different from most other human characteristics.

The strongly criterion-referenced procedure identifies a very large proportion of the candidates as having attainment below the lowest level available from each tier. Compared to operational GCSE examinations in Mathematics, where similar tiers are used, this implies that the teachers in the pilot study schools were singularly inept at entering their candidates for the right tier. The results of the conventional awarding procedure do not imply this.

Figure 8.1 shows in detail, for Tier 5 - 8, how the differences between the AT results from the conventional and strongly criterion-referenced awarding procedures arise. Similar data occur for the other tiers but these are not reported in detail because of the relatively small numbers of candidates involved. In Figure 8.1, the ranges of total scores within each Attainment Target are plotted separately for the candidates awarded each available level by the strongly criterion-referenced procedure. It can be seen that, although the candidates awarded different AT levels differ in terms of their mean total AT mark (and, in general but not always, those awarded higher levels have higher mean total AT marks) there is considerable overlap within each AT in the total AT marks scored by candidates awarded different levels. In particular, the ranges of marks scored by candidates who do not meet the two-thirds mastery score for any level and therefore are not awarded a level at all, extend almost up to half marks on all four Attainment Targets and overlap the range of marks scored by candidates awarded other

levels to a considerable extent. By definition, such overlaps cannot occur using conventional awarding procedures.

Figure 8.1
Ranges of total AT marks scored by candidates awarded each AT level by the strongly criterion-referenced procedure



It is important to be clear, however, that data such as these do not **necessarily** imply that the conventional awarding procedures are to be preferred to the strongly criterion-referenced ones. There is no external criterion variable available, against which to compare the two sets of Attainment Target results. All that can be said from a comparison of the results from the

two procedures, as applied to the Attainment Targets, is that they differ and that the strongly criterion-referenced ones have a number of counter-intuitive features. However, it is possible to go a little further by addressing the consistency of the internal workings of the strongly criterion-referenced procedure.

This procedure assumes that the questions on each Attainment Target are all correctly associated with a particular level. In addition, it assumes that the levels are ordered in a hierarchy of difficulty. If either of these assumptions does not hold, then there can be no sensible rationale for awarding each candidate the highest AT level for which their score exceeds the criterion mastery score. It was possible to test whether both of these assumptions held in the examination studied here and, in fact, one or both of them was frequently violated in the experimental examination. For each tier, Figure 8.2 plots the mean marks achieved by all the candidates on the questions associated with each level. If both assumptions of the strongly criterion-referenced procedure were to hold, the lines in these plots would all show a steady decline from top left to bottom right. Only AT4 shows this trend, and then only just, in all three plots.

8.9 THE WHOLE SUBJECT RESULTS FROM THE TWO PROCEDURES

As detailed in Sections 8.6 and 8.7, one of the key differences between the conventional and strongly criterion-referenced aggregation and awarding procedures lay in the way in which the Whole Subject levels were formed. Under the conventional procedures, the Whole Subject level was based upon the candidate's whole subject mark and a series of level thresholds; under the strongly criterion-referenced procedures, the Whole Subject level was the mean of the AT levels, rounded to the nearest integer.

The results of these two approaches are compared in this section. The teachers were asked to give an estimated Whole Subject level for each candidate, based upon their previous experience of GCSE examination standards. These, too, are reported here. Table 8.8 shows the distributions of the Whole Subject levels awarded by the two procedures and the teachers' estimates.

Figure 8.2
Mean percentage marks scored by all candidates entering each tier for questions associated with each level

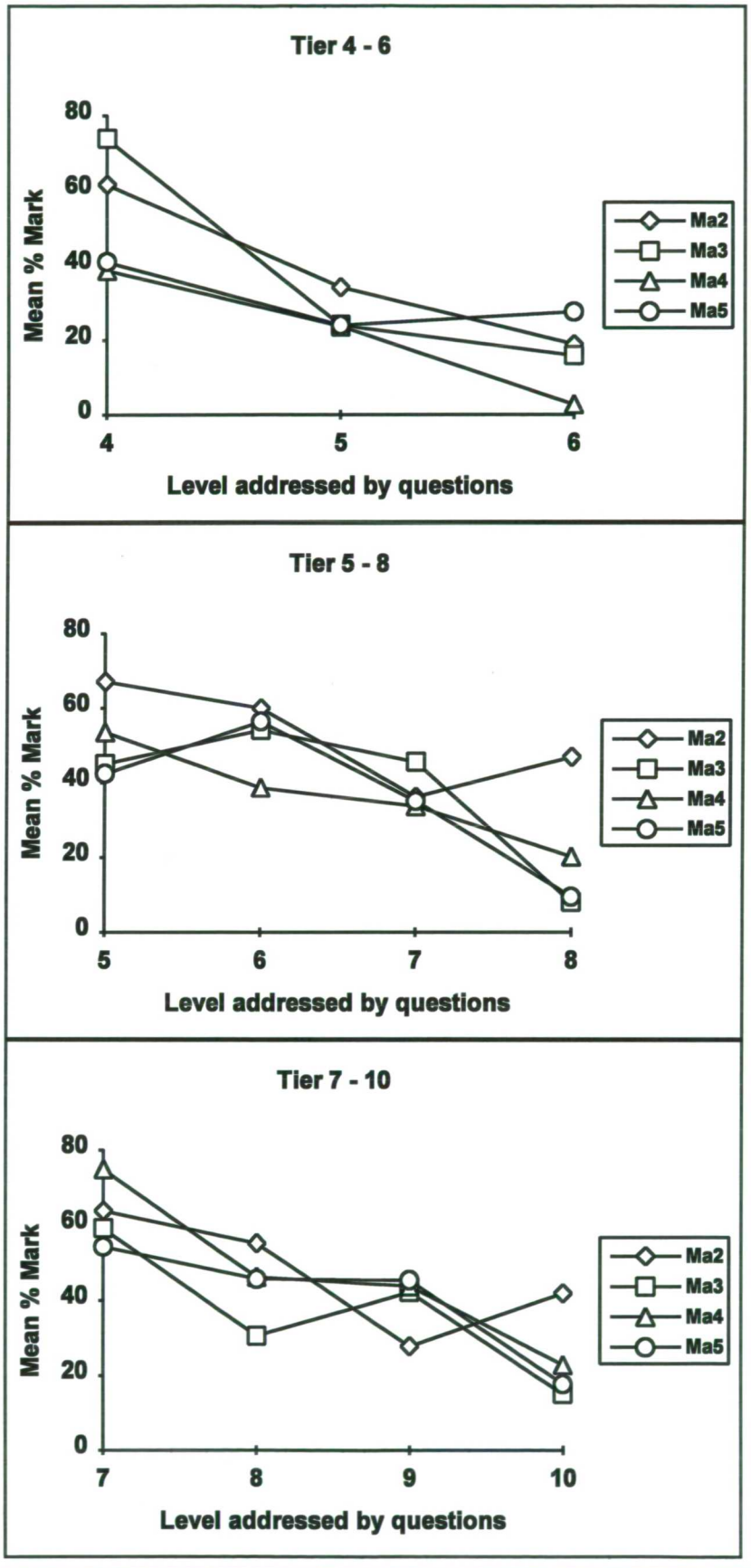


Table 8.8
Distributions of Whole Subject levels awarded by the conventional (Conv) and strongly criterion-referenced (S C-R) procedures, and teachers' estimates (Est)

Tier (no. of cands)	Level	Conv	S C-R	Est
4 - 6 (86)	U	22.1	83.7	0
	4	50.0	14.0	60.5
	5	26.7	2.3	26.7
	6	1.2	0	12.8
5 - 8 (275)	<5 = U	1.5	56.7	2.9
	5	29.8	9.1	10.9
	6	47.6	14.9	36.0
	7	17.1	17.1	46.2
	8	4.0	2.2	4.0
7 - 10 (128)	<7 = U	1.6	57.0	4.7
	7	37.5	14.8	35.9
	8	39.1	17.2	38.3
	9	18.8	10.9	11.7
	10	3.1	0	9.4

The principal difference between the distributions of Whole Subject levels awarded by the two procedures is the very high proportions of candidates who do not receive any level from the strongly criterion-referenced approach. As in the case of the AT levels, these proportions seem implausible, particularly in the light of the teachers' estimates. They reflect the high proportions of such results for the Attainment Targets (see Table 8.7).

They are also the result of the rules used to combine the AT results of candidates who have been not been awarded a level on one or more Attainment Targets. These rules were as follows:

1. If a candidate has not been awarded a level on one AT, this result is given a value of zero when the candidate's Whole Subject level (mean AT level) is computed.

2. If a candidate has not been awarded a level on two or more ATs, that candidate is not awarded a level for the Whole Subject.

These were the rules in use to derive whole subject results in Mathematics for National Curriculum assessments at Key Stage 3 at the time of the study. Their rationale was a strongly criterion-referenced one of wishing to ensure that candidates with a given whole subject level could be inferred to have achieved that level on each AT (hence, Rule 2), compromised by a recognition that perfection in this regard was asking for too much consistency from real candidates (hence, Rule 1). As the discussion in Chapter 4 showed, this sort of compromise greatly weakens the inferences about candidates' attainment which are the *raison d'etre* of strong criterion-referencing. However, Rule 1 is typical of the procedures used in practice when strong criterion-referencing is attempted. Such compromises are necessary because many candidates do not respond to real assessment instruments in a way which is consistent with the theoretical models of attainment underpinning strong criterion-referencing.

The differences between the Whole Subject levels awarded by the two procedures reflect the sensitivity of the strongly criterion-referenced procedure to small differences in where the candidates are successful. Table 8.9 illustrates this point with data for four Tier 5 - 8 candidates with the same total mark and, hence, the same conventionally awarded level.

Although these four candidates had identical total subject marks, and relatively small differences between their marks on the individual Attainment Targets, the differences between the Whole Subject levels awarded to them by the strongly criterion-referenced procedure are large. There are two main reasons for this. First, the effect of the method of determining the AT levels. This point is illustrated by Candidates 1 and 4 who scored the same total mark in AT5 but were awarded different levels for AT5 by the strongly criterion-referenced procedure because they scored their AT5 marks on different questions. The same effect also caused a difference of 5 marks in AT3 to change a Level 6 for Candidate 1 into no level at all for Candidate 4. The second reason for the observed differences in strongly criterion-referenced Whole Subject levels then came into play: the rules described above for combining the AT levels.

Table 8.9
Four candidates exemplifying the sensitivity of the strongly criterion-referenced procedure for awarding whole subject levels

Candidate	Total % Subject mark	Conv. Subject level	S C-R subject level	AT1 mark (S C-R level)	AT2 mark (S C-R level)	AT3 mark (S C-R level)	AT4 mark (S C-R level)	AT5 mark (S C-R level)
1	56	6	7	13 (7)	28 (8)	24 (6)	30 (7)	14 (6)
2	56	6	6	14 (7)	36 (8)	22 (5)	30 (8)	6 (U)
3	56	6	5	15 (7)	26 (8)	29 (6)	15 (U)	22 (6)
4	56	6	U	14 (7)	34 (8)	19 (U)	37 (8)	14 (U)

In Table 8.8 it was reported that the distributions of Whole Subject levels produced by the conventional awarding procedures are considerably more like the teachers' estimates than those from the strongly criterion-referenced procedure. However, such similarity could disguise substantial differences in the ordering of the candidates. Therefore, Table 8.10 shows the correlations between the Whole Subject levels awarded by the two procedures and the teachers' estimates. It can be seen that there is also better agreement in correlational terms between the conventionally awarded levels and the teachers' estimates. (The values in Table 8.10 are product-moment correlation coefficients computed after giving a value one less than the lowest available level of the tier to the U category of result. For example, in Tier 5 - 8, candidates to whom no level was awarded were notionally assigned to Level 4 for this analysis. Note also that the absolute values of the reported correlations are likely to be reduced by the small number of categories (4 or 5) on each variable. The data upon which Table 8.10 is based are given in full in Appendix 8.2)

Table 8.10
Correlation coefficients between Whole Subject levels awarded by the two procedures and teachers' estimated grades

Tier	Conventional	Criterion-referenced
4 - 6	0.60	0.54
5 - 8	0.69	0.61
7 - 10	0.80	0.68

8.10 A CASE STUDY OF ONE SCHOOL

This section reports, in more detail, the data from one of the schools which took part in the pilot examination. In line with their normal practice of entering some candidates early, this school also entered the study candidates for the normal SEG GCSE *Mathematics B* examination in November 1992. Table 8.11 shows the Whole Subject results obtained by this school in that operational GCSE examination and from the conventional and strongly criterion-referenced experimental examination procedures. The school in question only entered candidates for Tier 7 - 10.

Table 8.11
Whole Subject results from one school which also entered its candidates for a contemporaneous operational GCSE examination

Level	Number of candidates awarded level by:				
	Operational GCSE[1]	Teacher's estimates	Conventional procedure	S C-R procedure (2/3 mastery)	S C-R procedure (1/2 mastery)
10	4	5	4	0	5
9	16	11	16	11	22
8	12	15	11	12	4
7	0	1	1	5	1
U	0	0	0	4	0

[1] Converted from grades to levels for comparative purposes (see Table 8.5)

Particularly striking is that several of these candidates, who all got Levels 8 to 10 (Grades B or A) in the operational November examinations, were not awarded a level at all by the strongly

criterion-referenced procedures. An investigation was therefore made into what would have happened if the mastery scores in each level had been 50%. As can be seen from Table 8.11, this does, indeed, give every candidate a level but at the expense of awarding almost all the candidates Level 9. It is clear that, for this school at least, the experimental strongly criterion-referenced procedures did not produce results which were consistent with operational GCSE standards or with the teacher's views of the attainments of his own pupils.

Table 8.12 shows, however, that the conventional awarding procedure, applied to the experimental examination, did produce consistency with the operational results in this one school, at least.

Table 8.12
Cross-tabulation (numbers of candidates) of conventionally awarded levels from the experimental examination with operational GCSE results for one school.

		Conventional level from experimental exam.					
		U	7	8	9	10	Total
Operational level[1]	U						0
	7						0
	8		1	7	4		12
	9			4	11	1	16
	10				1	3	4
Total		0	1	11	16	4	32

[1] Converted from grades to levels for comparative purposes (see Table 8.5)

8.11 IN CONCLUSION

The comparison of the results of conventional and strongly criterion-referenced awarding procedures in this chapter has illustrated some of the practical effects of the theoretical concerns discussed in Chapter 3. These effects meant that the strongly criterion-referenced procedure did not produce results which were consistent with teachers' views of the attainment of their own pupils (or, indeed, with normal expectations about the distribution of attainment). The procedure studied was very closely modelled on one devised, for a slightly

different purpose, by committed proponents of strong criterion-referencing within the *Schools Examinations and Assessment Council* and among its technical advisers. However, the work reported here could be challenged by arguing that either a different criterion value would have produced better results or, more fundamentally, that different strongly criterion-referenced aggregation rules might have done so.

There are several answers to these challenges. First, the use of a different criterion value in the case of the school for which contemporaneous operational results were available did not produce more plausible results. Perhaps, with sufficient assiduity, one could have been found which would have done so, but this would seem to imply that criterion values must be determined, *post hoc*, by reference to the examination outcomes. Such an approach is not consistent with the intentions of strong criterion-referencing because it means that the specific inferences which can be drawn from candidates' reported results vary from assessment occasion to assessment occasion.

Different rules for aggregating the AT levels might, indeed, have produced Whole Subject results more consistent with the teachers' estimates. However, in order to do so, they would have had to compromise even further the specific inferences which could be drawn from candidates' reported Whole Subject results (see Section 8.9). This would represent a further step **away** from the purpose of strong criterion-referencing.

Finally, and most fundamentally, the data reported in Section 8.8 (Figure 8.2) showing the extent to which candidates' responses violated the theoretical difficulty hierarchy underpinning the strongly criterion-referenced procedures, imply that procedural changes like those considered above are irrelevant to the true problem. Although it is possible to argue that the data in Figure 8.2 reflect a failure of practice, rather than theory, and that other examiners might have been able to set papers which contained questions forming a hierarchy of difficulty according to the level with which they were associated, this seems very unlikely in the light of the relevant theoretical considerations and previous research discussed in Chapter 3 (Section 3.2.2).

In fact, the examiners involved in the present study were as experienced and well qualified as any and Macdonald (1993) reports that, after setting the question papers, they were clear that questions constructed to address the National Curriculum levels would not, thereby, exhibit increasing difficulty according to their associated level. Given this, even pre-testing questions and choosing those which are shown to fit the required difficulty hierarchy would be no solution since it would, perforce, lead to the use of questions which did not adequately represent the assessment domains defined for some levels. The strongly criterion-referenced procedure is therefore fundamentally inappropriate for the levels concerned.

Is the right conclusion, then, that the extensive development process, involving well-qualified national experts, which produced the National Curriculum levels in Mathematics failed to produce well-ordered hierarchical levels, but that a further attempt could be successful? Or is it more parsimonious to conclude, particularly in the light of all the other failures to achieve strong criterion-referencing, that the task of defining such levels is impossible? As pointed out in Chapter 3 (Section 3.2.2), the latter conclusion would be consistent with the accumulated evidence from many studies (reviewed by Cresswell and Houston, 1991) that the facility with which learnt knowledge and skills are used is **not** independent of the particular characteristics of the problems to which they are applied.

Context-free and well-ordered hierarchies of levels of skills and knowledge cannot, therefore, be constructed and, for this reason and for the technical reasons given in Chapter 4, the use of aggregation systems of even the most byzantine ingenuity will never enable valid specific inferences about candidates' attainment to be drawn from summary measures such as examination grades. Strong criterion-referencing is thus seen to be theoretically unsound and counter-intuitive results like those reported in this chapter are to be expected if it is attempted in practice.

CHAPTER 9

DISCUSSION, EVALUATION AND CONCLUSIONS

"This island was at a greater distance than I expected,
and I did not reach it in less than five hours."

A Voyage to Laputa from *Gulliver's Travels*
by Jonathan Swift

9.1 INTRODUCTION

In this concluding chapter, the key ideas and results from the preceding ones are drawn together. The necessary centrality, within conventional public examination awarding, of subjective evaluative judgements of quality was established in Chapter 3 (Section 3.4.2). In this chapter, a model of the cognitive process by which such judgements are made is proposed on the basis of the empirical work reported in Chapters 5 and 6. The alternative model which underpins the competing paradigm of strong criterion-referencing is identified and rejected on the basis of theoretical arguments developed in Chapter 3 (Section 3.3); its mismatch with the evaluative process in practice (Chapters 5 and 6); and the internal inconsistencies and improbable results of the experiment based upon it which was reported in Chapter 8.

An overall evaluation of conventional awarding procedures is given in this chapter, and a broadly positive conclusion reached. However, a number of significant problems and technical issues emerged during the work on these procedures which is reported in Chapters 5 and 6. These matters are summarised briefly in this chapter and, while keeping within the conventional paradigm, some major revisions to current awarding procedures are suggested in response. Further research is recommended to establish the feasibility of the suggested revised procedures.

First, however, the important practical question raised in Chapter 4 (Section 4.2) and investigated in the case study reported in Chapter 7 is addressed: how should standards set on each component of an examination be combined to determine candidates' grades for the examination as a whole?

9.2 COMBINING COMPONENT STANDARDS

9.2.1 Combining component grade boundary judgements

In Chapter 4 (Section 4.2.2) a number of models for combining grade boundary decisions on examination components were developed and discussed. However, a final consideration of their merits was delayed until this chapter so that it could be informed by the data reported in Chapters 5 and 6. The issue concerned the implicit question which awarders judging scripts are addressing. If they view each script in isolation, then it is appropriate to make allowance for effects in candidates' total marks due to regression to the mean when component boundaries are combined. On the other hand, if the awarders take account of the overall demands of the examination as a whole when considering each component, no such allowance is appropriate (see Section 4.2.2).

Since the present study has shown that awarders generally find contextualising their evaluative judgements difficult (see Sections 5.5 and 6.4, in particular), it must be concluded that combination methods which make allowance for regression effects should be used. In Section 4.2.2.5, the method of choice in this category was identified as:

$$B = \frac{\sum_{i}^{n}\left[(s_i^2 + \sum_{k\neq i}^{n} r_{ik}\cdot s_i\cdot s_k)\cdot F^{-1}\{f_i(b_i)\}\right]}{s_a^2} \qquad \text{Model 2b}$$

where B is the boundary mark on the total score scale of the examination as a whole, s_i is the standard deviation of the scaled scores on Component i, r_{ik} is the observed correlation between Components i and k, b_i is the corresponding raw score boundary for Component i, f_i is the cumulative distribution function for Component i, F is the cumulative distribution function for the examination's total score, s_a is the standard deviation of the aggregate scores and there are n components.

The strengths and weaknesses of this model were discussed in detail in Section 4.2.2.5 and that discussion will not be repeated here. However, further work is needed on the robustness of Model 2b when the number of candidates taking the examination is small. In particular,

it would be useful to investigate the use, in such circumstances, of the following model instead:

$$B = \frac{\sum_{i}^{n} \omega_i \cdot (b_i - m_{ri}) \cdot (s_i + \sum_{k \neq i}^{n} r_{ik} \cdot s_k)}{s_a} + m_a \qquad \text{Model 1b}$$

where ω_i is the scaling factor for Component i, m_{ri} is the raw (unscaled) mean for Component i and m_a is the mean aggregate score.

It is also worth mentioning that the anticipated transparency problems with models like Model 2b do, indeed, occur. In particular, Chapter 5 (Section 5.6.4) reports a tendency for awarders, having taken great care over their evaluative judgements on the separate components, to then abdicate responsibility for the final examination outcomes because they do not understand, and are sceptical about, the complex models, such as Model 2b, which are used to derive the aggregate examination boundaries from the component ones. This is both undesirable in itself and indicative of significant problems of transparency with such models.

9.2.2 Combining component grades

An alternative approach to combining component awarding decisions is to award a grade to each candidate on each component and then combine the component grades to produce an overall grade for the examination as a whole. Such an arrangement is transparent and has significant advantages in modular examinations, but at the price of some loss of reliability due to the combination of coarse component grade scales (see Chapter 4, Section 4.2.1). However, the work reported in Chapter 7 (Section 7.7) shows that the results of grade combination can be as acceptable in qualitative terms as the results of more conventional mark aggregation and awarding, provided that grade hurdles are not used. Further research into the use of grade aggregation is clearly desirable.

9.2.3 Avoiding the problem

One of the issues which the present study has not addressed is the possibility that awarders should make their judgements of candidates' work on the examination as a whole, rather than, as in current practice, on each component separately in turn. Clearly, this would completely avoid the problems associated with combining component judgements. There are significant practical difficulties, however, in modular examinations (where boundaries need to be set on some modules before the whole examination has been completed) and concerning coursework components (where assembling the complete work from several candidates at each aggregate mark point for judgemental purposes can be problematic and may be impossible because of the nature and bulk of the coursework). Moreover, awarders find it particularly difficult (see Chapter 5, Section 5.5.3) to evaluate what they call "unbalanced" scripts: those in which the quality of the answers varies greatly from question to question. It is to be anticipated that the scope and incidence of this problem would be significantly worsened if, say, two written papers and a coursework component had to be evaluated holistically. Nonetheless, useful further research could be done on the merits of using evaluative judgements in awarding based upon work from the examination as a whole, rather than each component in turn.

9.3 CURRENT CONVENTIONAL AWARDING PROCEDURES EXPLAINED AND EVALUATED

9.3.1 The nature of examination standards

It was established in Chapter 2 (Section 2.8) that, although they have many other roles and effects, the main function of public examinations has always been, and remains, the qualification of individuals for selection purposes. This has a number of significant consequences. Firstly, it is important on ethical grounds for the candidates whose life chances are affected by their examination results to be able to understand those results and, broadly, how they are determined. Secondly, examination results must be reported in a manner which is easily understood by, and useful to, the non-specialists who use them in subsequent selection processes. These two requirements for transparency impose substantial limitations on the complexity of examination procedures and reporting arrangements. In particular, they lead to the use of scales consisting of a small number of broad grades to report examination results (see Sections 2.8 and 2.9).

The interchangeable use, in practice, of results from many different examinations within the same selection process imposes the requirement that the grades awarded must be independent of any particular examination (within a given broad type of examinations, such as GCE A-levels). In other words, the standards represented by the grades must be comparable between different examination boards, between different assessments in the same subject from a given board and, indeed, between subjects themselves. Such comparability of standards is a necessary condition for fair selection under the meritocratic philosophy which underpins the use of examinations results in selection (see Section 2.8.2). In recent years, the role of examination results in published school performance tables and other monitoring activities has further increased demands for comparability between boards, between subjects and across time. (It is worth pointing out, however, that selection is normally between candidates who obtain their qualifications fairly contemporaneously. Some monitoring activities, on the other hand, imply the need for comparability over very much longer time periods, even though the culture-bound nature of examination standards (see Chapter 3, Section 3.4.2) renders the idea of comparability over long periods of time essentially meaningless.)

It was established in Chapter 3 (Section 3.4) that, for examination grades to be useful in selection, examination standards must reflect general social values about the status and utility of the knowledge and skills assessed. For example, the attainment required for a Grade C in GCSE English examinations must, broadly, reflect the expectations of society concerning the competence in this area of school leavers entering employment, whereas an A-level pass in English must require the attainment of knowledge and skills which are generally seen to reflect the desirability of, and the greater status accorded to, entry into Higher Education.

More fundamentally, the argument in Section 3.4 also showed that, to be philosophically coherent, any theory of standards which accomodates the notion of comparability across different assessment domains must define standards in terms of human value judgements. This is because quantitative comparisons of qualitatively different objects of any kind can have meaning only in terms of the value attached to the objects by human beings. The consequence of this is that setting examination grade standards is inevitably a value-laden, and essentially subjective, process of social construction.

Current public examination awarding procedures are consistent with this view. In these procedures, examination standards are created by special groups of judges, known as awarders, who are empowered, through the examining boards as government-regulated social institutions, to evaluate the quality of students' attainment on behalf of society as a whole. General acceptance of the results is encouraged by the use of a *due process* in the form of procedures which are nationally specified by a statutory authority.

9.3.2 The social process aspects of awarding

Defining the standards applied in a particular examination on a particular occasion by the consensus judgement of a **group** of awarders, rather than the only possible alternative - the judgement of a nominated individual, gives the process face validity and helps to protect it from idiosyncratic views. However, there will inevitably be social phenomena within such a group which affect its collective judgements. Chapters 5 and 6 report detailed data on these social effects as they operate in the conventional awarding process. Awarding meetings exhibit many of the well-established characteristics of small decision making groups. There is pressure, both informational and normative, on the participants to conform to the group's views, stereotyping of outgroup members and devaluing of information which is identified with an outgroup. This latter point is particularly important because it can apply to contextualising statistical information where a related social point concerns the reluctance of awarders sometimes to take appropriate cognisance of data which are at odds with their evaluative judgements. Given that the awarders are generally teachers whose professional skills involve judgement of their pupils' attainment, it is hardly surprising that they can feel reluctant to accept statistics which appear to cast doubt upon their professional competence but, when it occurs, this reluctance clearly works against the adequate contextualisation of awarding decisions.

In addition, because discussion in awarding meetings often takes the form of negotiation about the location of grade boundaries, rather than discussion of the evaluative worth of candidates' work, appeals to moral responsibility and similar rhetorical devices are common and personality factors clearly influence the decisions reached. As is to be expected, the role of

the Chair of awarding meetings is crucial. Also particularly worthy of mention is the way in which, during discussion of grade boundaries, **individual** cases are sometimes given greater weight than background statistical information based upon many hundreds of candidates' responses. This, too, is to be expected from the literature on human judgement (see, for example, Tversky and Kahneman, 1974). In the group setting of an award meeting, the persuasive power of individual cases of candidates whose work has been scrutinised sometimes contributes to a tendency to give the benefit of the doubt to candidates. While understandable, and even laudable, in human terms this tendency can produce improvements in examination outcomes which are difficult to justify in the light of the statistical information available. More rarely, the presence of a few candidates with high marks whose work is judged poor can have the reverse effect. In both cases, the *risky shift* appears to operate and the awarders as a group make decisions which are more extreme than most of them would make as individuals.

It is impossible not to conclude that many of the social factors which operate in awarding meetings work against the main purpose of the meetings: the setting of grade boundaries which make allowance for the difficulty of the current examination paper, and so represent a comparable standard to the grade boundaries used on a previous assessment occasion. Such social factors cannot be eradicated completely because of the necessary centrality of human judgements in standard setting but awarding procedures are needed which limit, as far as possible, their adverse effects. The work reported in Chapters 5 and 6 (Sections 5.6 and 6.5) shows that, in current procedures, these effects are constrained by appropriate definition of the relationships and roles of the participants in awarding meetings. Further research into the social psychology of conventional grade awarding meetings would be very useful.

9.3.3 The evaluative process in awarding - a possible cognitive model

In this section, a number of possible models of the cognitive process by which awarders make their evaluative judgements of the quality of candidates' work are briefly considered and the one which best fits the evidence gathered in this study is identified.

9.3.3.1 The Cartesian Computer model

This model sees awarders as identifying, in the work they are judging, the presence (or absence) of each element in an externally defined set of features (or *criteria*) which epitomise work of the standard required. They then apply a set of rules which tell them if sufficient of the features are present, in sufficient quantity, for the work to reach the standard required. The model is an essentially mechanical and dualist one, in which the awarders hold up the features of each candidate's work in some sort of mental *Cartesian theatre* (Dennett, 1993) and apply high level computational rules which determine its overall value.

Putting any scepticism about mind/brain dualism aside, there are still obvious problems with the Cartesian Computer model. Its key components, a fixed set of evaluatively relevant features which must be identified and computational rules to produce an evaluation from them, are the salient aspects of strong criterion-referencing. Indeed, proponents of strong criterion-referencing must be, perhaps unwittingly, subscribers to the Cartesian Computer model of evaluation. In Chapter 3 (Section 3.3) the epistemological and psychological naivety of this view of evaluation was discussed in depth. In particular, the notion that awarders' judgements involve the application to every piece of work of a fixed set of criteria, with fixed values, was contested. Moreover, the evidence given in Chapter 6 (Section 6.3) about awarders' reasons for their judgements does not support it. In Chapter 8 (Sections 8.8 to 8.10) the failure of Cartesian Computer mechanisms, when externalised in the aggregation rules of a strongly criterion-referenced examination, to produce results in line with normal evaluative judgements was reported. A model of evaluation other than the Cartesian Computer is clearly required.

9.3.3.2 The Cartesian Gestalt models

These models, like the Cartesian Computer model, are dualist but they do not presuppose a fixed set of computational rules. The awarders are still seen as first identifying the presence or absence of features of the work which are relevant to its evaluation. Once this has been done, an overall evaluation is produced of the evidence displayed in the Cartesian theatre, not by the application of fixed computational rules but by the direct perception of its gestalt. There are two variants of this model: in the *Fixed Cartesian Gestalt model*, the same supposedly evaluatively relevant features (or criteria) are applied to every piece of work evaluated and each such criterion is given the same value for any piece of work; in the *Variable Cartesian Gestalt model*, the evaluatively relevant features may differ from one piece of work to another

in terms of both relevance and value. The variable form of this model is made possible by the non-computational nature of the final evaluative step.

The Fixed Cartesian Gestalt model, like the Cartesian Computer, is inadequate because of its fixed set of evaluatively relevant features (see Sections 3.3.2 and 6.3.2). The Variable Cartesian Gestalt model, however, is broadly in line with conventional expectations about the awarding process (see, for example, Appendix 5.1). The conventional use, as a reference point, of archival work from a previous examination which has been awarded the grade being set assumes that this work can be compared (presumably in the Cartesian theatre) with the work being evaluated. The conventional notion that awarders should, and do, discuss the reasons for their evaluative judgements in order to arrive at a rational agreed compromise assumes that the process of evaluation is distinct from the identification of evaluatively relevant features. Finally, the idea that awarders can give an holistic characterisation of a piece of work which is worthy of a particular grade (a so-called *grade description*) again implies that the act of holistic evaluation is distinct from the identification of evaluatively relevant features.

The problem for the Variable Cartesian Gestalt model is that, as the evidence in Chapters 5 and 6 (Sections 5.5 and 6.4) overwhelmingly shows, awarders do not, left to their own devices, make properly contextualised *script as response* judgements. In addition, they only rarely, and then partially, discuss the reasons for their evaluations and when they are asked to do so they most frequently produce descriptions of their own affective response to the work rather than its evaluatively relevant features. Awarders' attempts to discuss the qualities of work which should be awarded a particular grade are generally brief, vague and partial. None of this is consistent with the Variable Cartesian Gestalt model's separation of the process of forming an evaluation from a prior process of identification and inspection of evaluatively relevant features of the work. If these features are on display in the Cartesian theatre, why do awarders find it so difficult to describe and discuss them?

9.3.3.3 The Zen model

The evidence from this study therefore suggests that a model of evaluation is required which does not separate the identification of evaluatively relevant features of candidates' work from

their evaluation. This suggests the Zen model which sees the awarders as perceiving the quality of the candidate's work directly and immediately, in the somewhat metaphysical way described by Pirsig (1974). This could also be called the *Water Beetle* model (see Chapter 1). The problem with this model is that evaluating a candidate's examination work is not, in general, an instantaneous process. The notion of immediate perception of overall quality and, thus, immediate holistic evaluation is therefore inadequate. The following model, however, corrects this deficiency of the plain Zen model.

9.3.3.4 The Multiple Zen Drafts model

This model sees the awarders as engaged in a constant process of evaluation and re-evaluation as they read the candidate's work. There is no need for a pre-existing set of evaluative criteria, although some loose and broad ones may exist, and therefore no set of computational rules for reaching an overall judgement. The evaluation is direct and immediate in the same way as for the plain Zen model but is continuously open to revision as the awarder reads more of the work. The awarder reads and re-reads the work until his or her evaluation stabilises. (See Dennett, 1993, for an extensive discussion of *multiple drafts* as a model of consciousness in general). The Multiple Zen Drafts model predicts that an awarder would have no difficulty in giving a current evaluation at any time during the process of reading a script (although such an evaluation might be accompanied by a caution to the effect that it could change when later material is read). This, they can, indeed, do.

The reader who is sceptical about the Multiple Zen Drafts model is invited to pause for an introspective moment and consider their own response to the present thesis. Is evaluation still being held off while an accurate description of the entire thesis is formed in a Cartesian theatre of the mind or is a view of the value of the work based upon the first eight chapters now being further modified during Chapter 9? Has reading this paragraph made absolutely no difference to your evaluation of the thesis! Has the incorrect punctuation mark at the end of the last sentence affected your evaluation? Could it have?

This last point illustrates the variable nature of evaluatively relevant criteria. In a competently punctuated piece of written work, the occasional error is generally judged insignificant but the presence of frequent punctuation errors may lead to punctuation becoming an evaluatively

relevant feature as the evaluator's judgement evolves. The example shows how the Multiple Zen Drafts model entails a mechanism which explains one of the defining characteristics (see Chapter 3, Section 3.3) of evaluative judgement: the nature of the work being evaluated *modifies the criteria relevant to its own evaluation.*

The data in Chapters 5 and 6 are most consistent with Multiple Zen Drafts as a cognitive model of evaluation, at least as far as public examination awarding judgements are concerned. In particular, this model explains the persistent failure of evaluative judgements of candidates' work to be *adequately contextualised* and the *lack of discussion of evaluatively* relevant features of candidates' work in awarding meetings. Taking the latter phenomenon first, the principal reason why awarders rarely offer each other detailed descriptions of the relevant features of candidates' scripts as reasons for their evaluations is that they normally do not form such descriptions at all; instead, they make an evolutionary succession of direct evaluations. As regards contextualisation, the occasional (Chapter 5, Section 5.5) references which awarders make to the question papers and archival material are inadequate because, according to the Multiple Zen drafts model, every step in the awarder's continuously evolving evaluation must use the current examination questions and the archival material as points of reference if the final judgement is to be fully contextualised. The Multiple Zen Drafts model also explains awarders' difficulty in judging "unbalanced" scripts (Chapter 5, Section 5.5.3). Such scripts will produce large changes between successive evaluations, making arriving at a stable judgement particularly difficult.

Thus, the process by which awarders judge candidates' work is one in which direct and immediate evaluations are formed and revised in an evolutionary way as the awarder reads and re-reads the work. At the conscious level, it is not a computational process and it cannot, therefore, be mechanised by the use of high-level rule-bound procedures and explicit criteria. Conventional awarding procedures are consistent with this cognitive model, strong criterion-referencing, the main competing paradigm, is its very antithesis.

There is considerable scope for further fundamental work on the way in which awarders make their evaluative judgements and on the effects of a range of variables upon the process. Much of the psychological work which has been done on physical and social judgement (see

Eiser, 1990, for a review) appears relevant to awarding judgements. In the present study there was evidence that awarders use an availability heuristic (Chapter 5, Section 5.5) and that their judgements are conditioned by local norms (Chapter 6, Section 6.6). Further systematic investigation of the extent of such effects and the incidence of other well-known effects, such as anchoring, would be well worthwhile. So, too, would studies of the reliability of awarders' evaluative judgements, perhaps in the context of experimental manipulations of the nature of the evidence with which they are presented.

9.3.4 Combining statistical and judgemental data to reach composite judgements

The evidence presented in Chapter 6 (Section 6.4) indicates that evaluative judgements of candidates work, by themselves, produce insufficiently contextualised decisions about the positions of grade boundaries. The use of such judgements as the main determiners of grade boundaries produces large annual changes of examination outcomes which, viewed statistically, occur with extremely improbable frequency. These can be called *Type 1* errors - changes in outcomes which are not justified. The Multiple Zen Drafts model of the way in which evaluative judgements are made suggests the reason for this result: the continuous evolution of awarders' evaluations does not enable contextualising information to be fully considered. Unfortunately, it was established in Chapter 3 (Section 3.5) that the maintenance of standards from one year to the next requires contextualised judgements of the *scripts as responses* to the question papers. This apparent impasse represents a major challenge to the conventional grade awarding paradigm which, necessarily, has evaluative judgements at its heart.

The solution to the problem which is currently used is to make the results of the qualitative judgements only one piece of information in a larger decision-making process which also includes explicit reference to contextualising statistical data. Chapter 6 (Section 6.6) shows that the use of statistical data to focus and modify awarders' qualitative evaluations of candidates' work can produce much more appropriately contextualised decisions, in the sense that the size and frequency of changes in examination outcomes become much more statistically plausible. The difficulty with current awarding procedures is knowing whether the statistical data are now so influential that they sometimes overwhelm the evaluative

judgements inappropriately, preventing changes in outcome when they should occur, and so creating *Type 2* errors. (The possibility of such Type 2 errors is, of course, the main argument against simply ensuring that the same proportion of candidates are awarded each grade on every assessment occasion.) For reasons discussed at length in Section 6.4, there is no way of knowing whether a Type 2 error has occurred in any particular examination but, based upon the work reported in Sections 6.4 and 6.6, Type 2 errors are likely to be considerably less frequent under current procedures than Type 1 errors are when statistical data are not used to provide an interpretative context for the awarders' evaluative judgements.

Current conventional awarding procedures are thus clearly an improvement upon previous ones since they produce more appropriately contextualised decisions than the previous use of awarders' evaluative judgements as the principal determiners of the positions of grade boundaries. However, some questions remain; for example, does the slight bias in the annual changes in outcomes which is visible in Figure 6.16 reflect a real improvement in candidates' attainment or an artefact of current awarding procedures? Therefore, while retaining the essential place of evaluative judgement as the basis for examination standards, the following section considers some alternative, more radical, possibilities for setting and maintaining comparable standards in public examinations.

9.4 POSSIBLE FUTURE DEVELOPMENTS WITHIN THE CONVENTIONAL PARADIGM

The first point which must be made in this section is to re-emphasise that examination standards must be based, at root, on human evaluative judgements of candidates' attainments (Chapter 3, Section 3.4). It is therefore essential for such judgements to have a central role in setting grade standards in the first examination on a new syllabus. However, one of the main messages of the present study (see Chapter 6, Sections 6.4 and 6.6) is that awarders find it extremely difficult to contextualise their evaluative judgements to take account of any changes there may have been in the difficulty of an examination between two successive assessment occasions. As a result, the central purpose of **annual** awarding meetings (as opposed to the first), which is to maintain comparable standards by making allowance for such changes, is threatened. The solution currently adopted to this problem is to guide the final decision-making process statistically. Awarding procedures therefore involve both qualitative and

quantitative data in a balance which is, itself, left to the judgement of those involved in each case.

Clearly, there is no difficulty in current procedures when the statistical data and evaluative judgements agree about the location of grade boundaries. The relative weights given to the two sources of information are then irrelevant to the final decision. However, a difficulty arises when, as is more common, the quantitative and qualitative data differ in their implications. In this case, the practical imperative (Chapter 2, Section 2.9) is relevant. It is not possible either to issue two sets of grades or to issue no grades at all while further evidence is gathered to settle the matter. A choice has to be made between the two sources of information or some compromise reached between them. This section looks at two alternative approaches to this problem.

9.4.1 Use a more formal Bayesian decision-making process

One superficially attractive way of addressing the problem would be to require the awarders to use a more formal approach, based upon Bayesian decision analysis (eg. Kaplan and Schwarz, 1975), to reconcile their evaluative judgements with the relevant statistical data. A more radical approach would be to relieve the awarders of the need to make final recommendations altogether and use them simply to provide indications of where the boundaries would lie, based only upon their evaluative judgements of candidates' work. These indications could then be combined with statistical data by others (perhaps the examining board's Chief Executive, who has the final approval of all awarding under current regulations - see SCAA 1994 and 1995) in a Bayesian decision-making process.

However, an explicit Bayesian approach would have several serious difficulties in the specific context of awarding (and see Collingridge, 1982, for a more general critique). The central problem would be the lack of prior probability or preference data. Bayesian approaches to reconciling statistical and judgemental indications of the positions of grade boundaries would require independent knowledge of the reliabilities of these two sources of information or, more directly, of the relative preference which should be given to them (see Upshaw, 1975; Rachlin, 1989). Such reliability information is unobtainable for the reasons given in Chapter 6: there is

no independent measure of the outcomes of any particular examination and so no criterion against which to judge the reliability of either source of evidence: evaluative judgements or statistical data. Although an alternative for the evaluative judgements might be to rely on the awarders' estimates of the certainty which they would attach to their own judgements, Nisbett and Wilson (1977) have shown how untrustworthy introspective access to degrees of certainty can be. As for preferences, Collingridge (1982) argues convincingly that people do not have privileged information about their own preferences.

Thus, to use a Bayesian approach in grade awarding, decision makers would have to decide, without evidence, upon the relative preferences they wish to give to the statistical and judgemental evidence. The effect would then simply be to produce final results like those implied by the judgements or those implied by the statistics or somewhere in between, depending upon the preferences used. The use of Bayesian decision theory would be essentially irrelevant to the result. Moreover, if a single set of preferences was agreed as part of the *due process* in nationally agreed procedures, a possible consequence might be that awarders would modify their reported qualitative judgements so that, when combined in the prescribed way with the statistical evidence, they produced a particular result. There thus seems little point, and some danger, in pursuing formal Bayesian approaches to the combination of qualitative and quantitative awarding data, unless adding a spurious air of scientific precision to the final positioning of grade boundaries is regarded as useful in itself.

9.4.2 Use only statistical data to maintain standards across years

Of course, Bayesian decision procedures of the type discussed in the preceding section simply provide a formal framework for reaching a compromise between the judgements and the statistics. To adopt informal case-by-case judgement to reach such a compromise, as is done in current procedures, does not avoid the basic problem: there is no evidence available with which to decide the relative weights to be given to the quantitative and qualitative data. A potentially more internally consistent alternative which should be considered, therefore, given the practical imperative always to produce a result, is simply to adopt one or other of the two types of data in all cases.

The evidence in Chapter 6 (Sections 6.4 and 6.6) unequivocally suggests that, if this were to be done, it should be the statistical, rather than the judgemental, data which is used to maintain standards. As noted earlier, the likely scale and incidence of Type 2 errors which would arise from using statistical data alone are much less than the likely scale and incidence of Type 1 errors created by the use only of evaluative judgement (see Section 6.4). It should also be noted that, with current procedures, the informal weighting of the two types of evidence means that either type of error can arise to an extent which cannot be known in any particular case.

A statistical appoach which could be used for maintaining standards was discussed in Chapter 3 (Section 3.5): the use of the *same schools definition* for the annual maintenance of grade awarding standards. As noted in Chapter 3, this can be seen as a more practical alternative to the use of the *same candidates definition* which, in the form of test equating, is used throughout the world. The test equating approach is not unproblematic, but the fundamental difficulties are inherent in any attempt to set comparable standards on two different assessment instruments. The need to do this is, of course, an unavoidable requirement of public examinations, however they are awarded, because of their selective purpose (Chapter 2, Section 2.8.2). Moreover, Goldstein (1986b) has argued that the strong technical requirements normally required for test equating can be relaxed for general comparability purposes allowing a sufficient, but more limited, notion of score *equivalence* to be used.

The advantages of this approach to the annual maintenance of standards would be considerable. First, in terms of year on year comparability, examination grades would be awarded in a transparent and objective way. The maintenance of standards would not be open to subjective challenge, either in general terms or by individual schools with a poor set of results in a particular year. Current awarding procedures render the outcomes of public examinations open to continual challenge in terms of credibility since they open up the possibility that the examination standards may have changed between successive assessment occasions. This is why public examination results, overall, can be greeted negatively whatever they may be. Either the outcomes in a particular year are worse, in which case it can be said that "standards of attainment in schools have fallen" or they are better in

which case it can be said that "examination standards have fallen". This charade is played out almost every year in the British media.

Nor would removing this potential challenge to the credibility of public examination results be an advantage only to examining boards. The annual press coverage is demoralising for both students and teachers because, however well they do, their success can be explained away as the result of lenient examining. It is also worth reflecting upon the possible effects which such credibility challenges have on the practice of selectors. To the extent that they sometimes appear to identify particular subjects or examining boards as deviant, challenges to the maintenance of standards may well lead to substantial injustice in selection procedures.

Second, statistically maintained awarding standards on unchanged syllabuses would give examination results which were more stable over short time periods and the results would be more useful for selection as a result. In particular, the erroneous extreme changes in outcomes (Type 1 errors) sometimes produced by awarders' judgements would not occur. (Note that if current procedures are operated in such a way as to prevent such changes ever affecting candidates' results, this is tantamount to the proposed use of only statistical data, but without the major transparency advantages just discussed.)

The third advantage of the proposed approach to the annual maintenance of standards would be that it would free resources, of both personnel and time, within examining boards. Such resources could then be used more intensively in awarding grades for the first time on examinations with new syllabuses (see Section 9.4.3). Some of the released resources would also be available for other purposes; in particular, improvements in marking reliability might become possible.

The main argument likely to be raised against the use of the *same schools definition* for the annual maintenance of standards is that, because some Type 2 errors would occur, it would make impossible the use of public examination results for the long term monitoring of the educational system. This follows because any overall changes in attainment in the centres common to two adjacent assessment occasions would effectively be removed from the results. In reality, however, the loss of this function is only the loss of a chimera since, as

Newton (1996) recently demonstrated so convincingly, examination results, however they are awarded, cannot validly be used as indicators of changes in educational standards over time. Moreover, the occurrence of Type 2 errors, and the size of any overall changes consequentially being removed from the results, could be investigated by periodic research using the *same candidates definition*.

An alternative, but similar, approach to the annual maintenance of grade standards would be the use of the *catch-all definition* of comparability in the same way (see Chapter 3, Section 3.4.1.6). This would require the routine collection of more information about candidates and their schools than is presently done but would be technically preferable to the *same schools definition*. Such an approach would build on the work begun by Nuttall and Armitage (1984).

Further work on replacing the use of statistically informed evaluative judgements with purely statistical approaches for the annual maintenance of standards in successive examinations on an unchanging syllabus should be done urgently. This work should address, *inter alia*, the robustness of the various possible approaches, particularly for examinations with relatively small numbers of candidates.

9.4.3 For the first examination on a new syllabus, retain current practice but involve participants representing a wider range of interests

Even if the annual maintenance of standards is done using a statistical definition of comparability, evaluative judgements **must** be involved when grades are awarded for the first examination on a new syllabus (see Chapter 3, Section 3.4). The question then arises of whether current practice is the optimum approach to use for this purpose. Lindblom (1965) and Collingridge (1982) have written persuasively that socially important decisions often are, and always should be, made by a process of reasoned argument between parties representing different groups with legitimate interests in the outcome. To the extent that current awarding procedures represent such a reasoned argument between awarders who make professional judgements and board officers with statistical concerns, they follow this model. Christie and Forrest (1981) argued similarly when they characterised decision making in awarding meetings as a benign "contest" between awarders and board staff.

However, it is possible to argue that awarders and examining board staff do not represent all the parties legitimately interested in the outcomes of the awarding process. Others who may have a legitimate interest include candidates, parents, teachers, selectors and statutory regulatory bodies such as the *School Curriculum and Assessment Authority*. As set out in Chapter 3, the definition of comparable standards depends upon trust. Until recently, examining boards were trusted to represent the views of all interested parties when they were setting standards. However, a significant departure from this has recently (1994) been introduced by the regulators. The *School Curriculum and Assessment Authority* has appointed *assessors* who attend the awarding meetings of each examining board in certain subjects and exercise considerable influence over the decisions made. In terms of the analysis in Chapter 3, this could be presented as an attempt to ensure that examination standards better reflect the values of society as a whole.

There is, therefore, a case for broadening the range of interests involved in setting grade standards still further. From the examining boards' perspective, this might help to counter recent arguments that they are partly motivated by self interest when grades are awarded, although, on the other hand, the involvement of Governmental bodies in the process also raises potential questions about **political** self interest. The inclusion of awarders from all the examining boards in such a group might also have potential for improving the comparability of standards between boards. Indeed, if the logic of the *social definition* of comparability (Chapter 3, Section 3.4.2) is followed through to its conclusion, the decisions of a **national** awarding committee representing all interested parties and using a *due process* of **nationally** agreed procedures would be theoretically unchallengeable (see Newton, 1996) although the serious equity issues discussed by Goldstein and Cresswell (1996) would arise if they produced sudden large changes in standards.

Significant practical difficulties surrounding the availability of individuals would arise if the range of participants in **annual** awarding meetings were to be greatly extended in the way suggested above. However, if the annual maintenance of standards were to be done statistically, as suggested in the preceding section, it would be more practicable to use extended groups of awarders to set standards on the first examinations of new or substantially revised syllabuses. The extension of representation at the first awarding meeting on new

syllabuses should therefore be given further active consideration which will need to cover three main aspects: the question of exactly which interests should be represented, and by whom; how to decide when a revised syllabus is sufficiently new to require new qualitative judgements to be made; and the awarding procedures which should form the *due process* of the enlarged, and possibly less specialist, awarding meetings involved.

9.5 IN CONCLUSION

This study had the following aims (see Chapter 1):

1. to develop a better theoretical basis for public examination awarding;
2. to investigate the nature of the qualitative judgements upon which it depends;
3. to investigate the information used in current awarding procedures and the way in which it is used;
4. to explore possible alternative approaches to awarding public examination grades;
5. to evaluate, on the basis of 1 to 4, current awarding procedures and, if appropriate,
6. to recommend changes to current procedures.

In respect of Aim 1, the fundamentally evaluative nature of examination awarding has been elucidated, allowing the problem of defining comparability of standards to be addressed from a new perspective. A new definition of comparability - the *Social Value Definition* - has been proposed which, for the first time, provides a formally satisfactory basis for defining comparable standards across different assessment domains. Examination standards have consequentially been identified for what they are - social constructions - and awarding has been shown to be **essentially** subjective in nature.

As far as Aim 2 is concerned, the evaluative process has been identified as one of continual revision of immediate perceptions of the value of candidates' work - this approach has been dubbed the *Multiple Zen Drafts Model*. Observations and analysis of awarders making judgements have shown that the models of awarders' qualitative judgements usually thought to describe the awarding process - the *Variable or Fixed Cartesian Gestalt Models* - are not accurate. The *Multiple Zen Drafts Model* helps to explain why attempts to reproduce

awarders' qualitative judgements with the use of explicit criteria and high-level computational rules must fail.

Aim 4 mainly related to the evaluation of just such an attempt: the principal competing paradigm to conventional grade awarding - *Strong Criterion-Referencing.* The theoretical basis of this was shown to be näive and to make use of a model of the evaluative process - the *Cartesian Computer Model* - which does not fit the observed behaviour of awarders. Nonetheless, data from a substantial experimental attempt to use strong criterion-referencing were analysed in Chapter 8 and the experimental assessments were found not only to produce the implausible results predicted by theory but also to suffer from significant internal inconsistencies.

Although not originally conceived as such, the present study took on some of the characteristics of *Action Research.* In keeping with Aims 3, 5 and 6, the results of the first two phases of the work were the main impetus behind some major changes in awarding procedures. During Phases 1 and 2 of the study, the awarders' evaluative judgements of candidates' work were identified as the principal data used for awarding. In most cases, these judgements alone determined the final examination outcomes. Statistical data were referred to briefly and only rarely taken into account in the final positioning of grade boundaries. A number of adverse consequences of the social nature of the awarding meetings as small decision-making groups were also identified in Phases 1 and 2. These exacerbated the tendency of awarders to give little attention to the statistical evidence. The procedures in use during Phases 1 and 2 of the study produced improbably frequent and large changes in examination outcomes from one year to the next.

The main procedural changes made following Phases 1 and 2 of the study were intended to ensure that statistical data exerted more influence on the awarding decisions and were considered alongside the results of the awarders' evaluative judgements. Other changes were made to alleviate some of the worst negative effects arising from social phenomena within the awarding meetings. These changes were evaluated during Phase 3 and found to be largely successful in their effects. The nature and scale of annual changes in outcomes were shown to have been brought more into line with expectations based upon sampling

theory and what little is known about the speed and prevalence of large-scale educational change. The roles and responsibilities of the participants in awarding meetings became more distinct, allowing a more even-handed consideration of the different evidence which different participants brought to the meetings. In particular, a more even balance was shown to have been struck between the qualitative and statistical data.

Aim 6 was thus met during the course of the study, following Phase 2. In addition, significant recommendations for further work on possible new procedures to overcome some of the remaining difficulties have been made in this chapter, following the evaluation of Phase 3. Several avenues for further research have also been identified.

9.6 POSTSCRIPT

Throughout the writing of this thesis, it has been impossible not to be aware of the irony of constructing written work for evaluation which analyses, evaluates and significantly critiques the very process for which it is destined. As Douglas Hofstadter (1980) has so elegantly illustrated, self reference is frequently a creator of paradox and it has almost seemed possible at times that a physical version of the Epimenides, or liar, paradox might be created. Had the conclusion of this study been that the evaluative judgements made in examination awarding were flawed beyond repair, any examiner finding the thesis persuasive would, thereby, have been rendered unable honestly to examine it. The negative consequences for the award are all too obvious.

Fortunately, the conclusion is not that examination grades cannot be awarded but, rather, that to do so requires well designed procedures, based on a proper understanding of the socially constructed nature of educational standards, which ensure that the results of the essentially subjective value judgements involved are adequately contextualised. From this starting point, there is scope for much further work on both the fundamental nature of the judgemental process and on practical procedures:

> "And to make an end is to make a beginning.
> The end is where we start from."
>
> *Little Gidding* - T S Eliot

APPENDIX 4.1

DATA FOR TABLE 4.2

		Biology				Business Studies			
		stats	A	B	E	stats	A	B	E
Paper 1	Scaling factor	1				1			
	Intended weight	38.5				50.0			
	Achieved weight	44.2%				45.2%			
	Scaled mean	44.8				60.0			
	Scaled s.d.	15.0				16.1			
	Grade boundary		64	56	35		87	72	50
	Cumulative % at bdy		13.20	26.15	73.87		5.41	22.80	74.42
Paper 2	Scaling factor	1				1.5			
	Intended weight	46.2				50.0			
	Achieved weight	51.6%				55.0%			
	Scaled mean	45.0				60.9			
	Scaled s.d.	17.3				18.8			
	Grade boundary		72	58	32		61	54	35
	Cumulative % at bdy		8.69	25.56	76.71		6.10	15.19	68.10
Paper 3	Scaling factor	1.33							
	Intended weight	15.4							
	Achieved weight	4.4%							
	Scaled mean	36.7							
	Scaled s.d.	3.3							
	Grade boundary		30	28	24				
	Cumulative % at bdy		19.98	62.79	93.29				
Paper 4	Scaling factor								
	Intended weight								
	Achieved weight								
	Scaled mean								
	Scaled s.d.								
	Grade boundary								
	Cumulative % at bdy								
Paper 5	Scaling factor								
	Intended weight								
	Achieved weight								
	Scaled mean								
	Scaled s.d.								
	Grade boundary								
	Cumulative % at bdy								
Inter-component correlations	r_12	0.86				0.58			
	r_13	0.33							
	r_14								
	r_15								
	r_23	0.34							
	r_24								
	r_25								
	r_34								
	r_35								
	r_45								
Aggregate Statistics	Mean	126.6				121.1			
	s.d	32.4				31.0			
Aggregate mark with same cumulative % as on paper:	Paper 1		167	150	104		172	144	102
	Paper 2		177	151	102		170	154	108
	Paper 3		158	114	81				
	Paper 4								
	Paper 5								
Aggregate grade boundaries	Model 1		175.9	151.2	98.9		178.5	153.0	102.5
	Model 1b		172.4	150.1	102.6		172.4	149.8	104.8
	Model 2		169.0	144.0	101.0		171.0	149.0	105.0
	Model 2a		170.2	144.9	99.5		171.0	149.0	105.0
	Model 2b		171.7	148.9	102.0		170.9	149.5	105.3

		Communication Studies				Economics			
		stats	A	B	E	stats	A	B	E
Paper 1	Scaling factor	2				1			
	Intended weight	40.0				25.0			
	Achieved weight	41.1%				23.5%			
	Scaled mean	85.0				26.1			
	Scaled s.d.	25.8				8.0			
	Grade boundary		63	55	38		39	33	22
	Cumulative % at bdy		7.04	17.80	63.43		7.90	23.27	68.44
Paper 2	Scaling factor	1.5				1			
	Intended weight	30.0				25.0			
	Achieved weight	27.1%				21.0%			
	Scaled mean	72.7				20.4			
	Scaled s.d.	19.4				7.5			
	Grade boundary		66	58	37		35	26	16
	Cumulative % at bdy		9.18	23.91	82.58		4.79	24.03	73.25
Paper 3	Scaling factor	1				1			
	Intended weight	30.0				50.0			
	Achieved weight	31.8%				55.3%			
	Scaled mean	85.3				41.1			
	Scaled s.d.	22.0				16.5			
	Grade boundary		105	85	68		64	55	34
	Cumulative % at bdy		20.02	53.27	79.32		9.86	21.59	66.09
Paper 4	Scaling factor								
	Intended weight								
	Achieved weight								
	Scaled mean								
	Scaled s.d.								
	Grade boundary								
	Cumulative % at bdy								
Paper 5	Scaling factor								
	Intended weight								
	Achieved weight								
	Scaled mean								
	Scaled s.d.								
	Grade boundary								
	Cumulative % at bdy								
Inter-component correlations	r_12	0.38				0.57			
	r_13	0.36				0.63			
	r_14								
	r_15								
	r_23	0.31				0.58			
	r_24								
	r_25								
	r_34								
	r_35								
	r_45								
Aggregate Statistics	Mean	243.3				87.6			
	s.d	50.8				27.8			
Aggregate mark with same cumulative % as on paper	Paper 1		320	290	227		130	109	73
	Paper 2		312	279	198		138	108	69
	Paper 3		285	239	204		127	111	75
	Paper 4								
	Paper 5								
Aggregate grade boundaries	Model 1		330.0	282.0	199.5		138.0	114.0	72.0
	Model 1b		309.6	273.4	211.1		130.8	110.5	74.2
	Model 2		305.0	269.0	212.0		129.0	110.0	73.0
	Model 2a		307.1	271.4	211.4		130.5	109.8	73.0
	Model 2b		306.7	270.8	211.8		130.0	109.9	73.3

		English I				English III			
		stats	A	B	E	stats	A	B	E
Paper 1	Scaling factor	2				1			
	Intended weight	33.3				22.2			
	Achieved weight	35.2%				22.6%			
	Scaled mean	51.7				39.9			
	Scaled s.d.	16.2				14.0			
	Grade boundary		37	32	20		56	51	29
	Cumulative % at bdy		9.81	24.57	77.68		15.15	22.89	78.49
Paper 2	Scaling factor	1				1.25			
	Intended weight	33.3				27.8			
	Achieved weight	30.7%				33.0%			
	Scaled mean	52.3				52.2			
	Scaled s.d.	13.7				18.0			
	Grade boundary		72	63	41		57	52	25
	Cumulative % at bdy		8.53	22.93	80.65		18.04	25.81	88.91
Paper 3	Scaling factor	1				1			
	Intended weight	33.3				50.0			
	Achieved weight	34.5%				44.1%			
	Scaled mean	52.1				117.6			
	Scaled s.d.	15.2				22.5			
	Grade boundary		74	64	42		145	126	85
	Cumulative % at bdy		8.94	21.84	76.31		14.25	35.87	93.62
Paper 4	Scaling factor								
	Intended weight								
	Achieved weight								
	Scaled mean								
	Scaled s.d.								
	Grade boundary								
	Cumulative % at bdy								
Paper 5	Scaling factor								
	Intended weight								
	Achieved weight								
	Scaled mean								
	Scaled s.d.								
	Grade boundary								
	Cumulative % at bdy								
Inter-component correlations	r_12	0.42				0.42			
	r_13	0.4				0.46			
	r_14								
	r_15								
	r_23	0.55				0.55			
	r_24								
	r_25								
	r_34								
	r_35								
	r_45								
Aggregate Statistics	Mean	156.1				209.8			
	s.d	35.9				44.5			
Aggregate mark with same cumulative % as on paper	Paper 1		204	182	130		260	245	174
	Paper 2		207	184	126		255	241	158
	Paper 3		206	185	132		263	225	147
	Paper 4								
	Paper 5								
Aggregate grade boundaries	Model 1		220.0	191.0	123.0		272.3	242.0	145.3
	Model 1b		207.2	184.0	129.7		260.8	235.5	156.5
	Model 2		206.0	183.0	129.0		260.0	234.0	158.0
	Model 2a		205.7	183.7	129.3		260.1	233.9	156.1
	Model 2b		205.6	183.6	129.5		259.7	234.8	156.8

		French				Physics A			
		stats	A	B	E	stats	A	B	E
Paper 1	Scaling factor	1.2				3.5			
	Intended weight	20.0				35.0			
	Achieved weight	19.3%				35.5%			
	Scaled mean	45.0				95.1			
	Scaled s.d.	15.0				30.0			
	Grade boundary		47	38	24		37	32	22
	Cumulative % at bdy		23.07	46.58	86.53		17.17	32.08	71.14
Paper 2	Scaling factor	1				1.88			
	Intended weight	20.0				45.0			
	Achieved weight	19.3%				50.3%			
	Scaled mean	52.4				123.0			
	Scaled s.d.	14.8				40.3			
	Grade boundary		67	58	39		92	85	57
	Cumulative % at bdy		18.96	37.63	81.64		13.13	22.38	66.92
Paper 3	Scaling factor	1				1.11			
	Intended weight	20.0				20.0			
	Achieved weight	21.5%				14.3%			
	Scaled mean	40.5				69.5			
	Scaled s.d.	16.3				15.0			
	Grade boundary		61	53	34		73	66	53
	Cumulative % at bdy		12.16	23.77	64.31		25.13	48.43	79.92
Paper 4	Scaling factor	1							
	Intended weight	20.0							
	Achieved weight	20.3%							
	Scaled mean	39.9							
	Scaled s.d.	18.1							
	Grade boundary		62	51	31				
	Cumulative % at bdy		11.95	26.19	68.08				
Paper 5	Scaling factor	1.33							
	Intended weight	20.0							
	Achieved weight	19.9%							
	Scaled mean	56.1							
	Scaled s.d.	16.5							
	Grade boundary		54	46	32				
	Cumulative % at bdy		17.77	40.21	81.11				
Inter-component correlations	r_12	0.7				0.81			
	r_13	0.63				0.52			
	r_14	0.45							
	r_15	0.61							
	r_23	0.72				0.65			
	r_24	0.49							
	r_25	0.54							
	r_34	0.5							
	r_35	0.59							
	r_45	0.4							
Aggregate Statistics	Mean	234.1				287.8			
	s.d	64.7				77.2			
Aggregate mark with same cumulative % as on paper	Paper 1		283	239	161		370	332	244
	Paper 2		293	253	175		383	357	253
	Paper 3		312	282	209		349	295	221
	Paper 4		312	276	202				
	Paper 5		295	249	176				
Aggregate grade boundaries	Model 1		318.2	268.8	175.4		383.5	345.1	243.0
	Model 1b		302.5	262.4	186.4		375.8	341.4	248.2
	Model 2		298.0	260.0	186.0		371.0	335.0	244.0
	Model 2a		299.0	259.8	184.6		371.7	335.9	243.5
	Model 2b		299.4	260.4	185.2		373.5	339.3	245.2

		Physics B				Psychology			
		stats	A	B	E	stats	A	B	E
Paper 1	Scaling factor	3.5				2			
	Intended weight	35.0				40.0			
	Achieved weight	35.0%				39.9%			
	Scaled mean	92.6				87.1			
	Scaled s.d.	29.6				31.7			
	Grade boundary		37	32	22		68	58	37
	Cumulative % at bdy		14.64	27.91	69.26		6.44	19.09	66.64
Paper 2	Scaling factor	1.88				2			
	Intended weight	45.0				40.0			
	Achieved weight	51.6%				42.2%			
	Scaled mean	116.2				85.8			
	Scaled s.d.	41.4				33.1			
	Grade boundary		92	85	57		65	56	35
	Cumulative % at bdy		10.35	17.36	61.68		10.36	23.45	67.46
Paper 3	Scaling factor	1.11				1			
	Intended weight	20.0				20.0			
	Achieved weight	13.7%				17.9%			
	Scaled mean	72.4				69.2			
	Scaled s.d.	14.3				18.0			
	Grade boundary		75	68	57		87	78	47
	Cumulative % at bdy		26.24	49.14	78.25		16.59	35.38	89.26
Paper 4	Scaling factor								
	Intended weight								
	Achieved weight								
	Scaled mean								
	Scaled s.d.								
	Grade boundary								
	Cumulative % at bdy								
Paper 5	Scaling factor								
	Intended weight								
	Achieved weight								
	Scaled mean								
	Scaled s.d.								
	Grade boundary								
	Cumulative % at bdy								
Inter-component correlations	r_12	0.81				0.66			
	r_13	0.54				0.47			
	r_14								
	r_15								
	r_23	0.65				0.49			
	r_24								
	r_25								
	r_34								
	r_35								
	r_45								
Aggregate Statistics	Mean	281.6				242.1			
	s.d	77.4				70.2			
Aggregate mark with same cumulative % as on paper	Paper 1		372	333	243		348	306	212
	Paper 2		387	363	261		332	296	210
	Paper 3		337	288	220		312	271	153
	Paper 4								
	Paper 5								
Aggregate grade boundaries	Model 1		385.7	347.3	247.4		353.0	306.0	191.0
	Model 1b		378.2	343.7	251.9		337.3	296.3	200.0
	Model 2		371.0	335.0	248.0		333.0	295.0	202.0
	Model 2a		371.8	337.5	246.5		334.4	295.0	199.4
	Model 2b		375.0	342.3	249.1		334.8	295.5	200.6

		Pure & Applied Maths				Sociology I			
		stats	A	B	E	stats	A	B	E
Paper 1	Scaling factor			1	1	1.33			
	Intended weight	1.0		50.0	50.0	50.0			
	Achieved weight	56.7%				44.6%			
	Scaled mean	50.0		57.4	57.4	44.7			
	Scaled s.d.	57.4		26.9	26.9	13.0			
	Grade boundary	26.9	88	77	46		48	42	30
	Cumulative % at bdy		15.92	29.86	67.02		7.68	20.25	65.39
Paper 2	Scaling factor	1		1	1	1			
	Intended weight	50.0		50.0	50.0	50.0			
	Achieved weight	43.6%				56.0%			
	Scaled mean	35.4		35.4	35.4	40.8			
	Scaled s.d.	21.1		21.1	21.1	15.7			
	Grade boundary		64	49	25		62	55	37
	Cumulative % at bdy		11.98	26.05	64.38		8.60	19.08	61.67
Paper 3	Scaling factor								
	Intended weight								
	Achieved weight								
	Scaled mean								
	Scaled s.d.								
	Grade boundary								
	Cumulative % at bdy								
Paper 4	Scaling factor								
	Intended weight								
	Achieved weight								
	Scaled mean								
	Scaled s.d.								
	Grade boundary								
	Cumulative % at bdy								
Paper 5	Scaling factor								
	Intended weight								
	Achieved weight								
	Scaled mean								
	Scaled s.d.								
	Grade boundary								
	Cumulative % at bdy								
Inter-component correlations	r_12	0.84				0.66			
	r_13								
	r_14								
	r_15								
	r_23								
	r_24								
	r_25								
	r_34								
	r_35								
	r_45								
Aggregate Statistics	Mean	92.4				85.5			
	s.d	46.0				26.1			
Aggregate mark with same cumulative % as on paper	Paper 1		144	122	70		122	108	76
	Paper 2		153	127	74		121	109	79
	Paper 3								
	Paper 4								
	Paper 5								
Aggregate grade boundaries	Model 1		152.0	126.0	71.0		125.8	110.9	76.9
	Model 1b		149.3	124.3	71.5		122.4	108.7	77.7
	Model 2		148.0	124.0	72.0		121.0	109.0	78.0
	Model 2a		148.5	124.5	72.0		121.5	108.5	77.5
	Model 2b		147.9	124.2	71.7		121.4	108.6	77.7

		Sociology II				Theatre Studies			
		stats	A	B	E	stats	A	B	E
Paper 1	Scaling factor	1.33				1.25			
	Intended weight	50.0				35.0			
	Achieved weight	48.3%				30.2%			
	Scaled mean	46.7				57.0			
	Scaled s.d.	12.6				12.4			
	Grade boundary		48	42	30		55	49	32
	Cumulative % at bdy		9.53	24.05	71.76		19.86	40.83	91.67
Paper 2	Scaling factor	1				1			
	Intended weight	25.0				40.0			
	Achieved weight	24.0%				41.8%			
	Scaled mean	22.2				46.6			
	Scaled s.d.	7.2				14.7			
	Grade boundary		32	26	18		72	62	42
	Cumulative % at bdy		8.90	32.73	74.55		4.90	15.37	63.27
Paper 3	Scaling factor	1				1.25			
	Intended weight	25.0				25.0			
	Achieved weight	27.5%				27.9%			
	Scaled mean	32.0				29.6			
	Scaled s.d.	8.4				10.8			
	Grade boundary		42	35	28		40	33	21
	Cumulative % at bdy		12.60	41.64	70.69		4.77	16.60	62.11
Paper 4	Scaling factor								
	Intended weight								
	Achieved weight								
	Scaled mean								
	Scaled s.d.								
	Grade boundary								
	Cumulative % at bdy								
Paper 5	Scaling factor								
	Intended weight								
	Achieved weight								
	Scaled mean								
	Scaled s.d.								
	Grade boundary								
	Cumulative % at bdy								
Inter-component correlations	r_12	0.57				0.33			
	r_13	0.49				0.31			
	r_14								
	r_15								
	r_23	0.44				0.49			
	r_24								
	r_25								
	r_34								
	r_35								
	r_45								
Aggregate Statistics	Mean	100.9				133.4			
	s.d	23.3				29.1			
Aggregate mark with same cumulative % as on paper	Paper 1		132	118	88		159	140	94
	Paper 2		133	112	87		182	164	124
	Paper 3		128	106	89		183	163	125
	Paper 4								
	Paper 5								
Aggregate grade boundaries	Model 1		137.8	116.9	85.9		190.8	164.5	108.3
	Model 1b		131.4	114.3	88.5		178.1	157.9	115.0
	Model 2		131.0	114.0	88.0		171.0	154.0	116.0
	Model 2a		131.3	113.5	88.0		174.2	155.4	113.8
	Model 2b		131.1	113.3	88.0		175.3	156.5	115.2

APPENDIX 5.1

AWARDING PROCEDURE DOCUMENT IN USE DURING PHASES 1 AND 2

GUIDELINES

for the conduct of

GCE ADVANCED AND ADVANCED SUPPLEMENTARY LEVEL GRADE AWARDING MEETINGS

This document provides guidelines covering the procedure for determining grade boundaries and does not include such matters as possible training for grade awarders, aegrotat awards, cheating cases or other matters to do with assigning or issuing grades.

The revisions in this issue are minor and mainly involve additions of references to 'AS' Level and other editorial amendments. For this reason, the revisions have not been side-lined.

Introduction

(a) The determination of grade boundaries is a crucial part of the examination process. The process begins with the setting of question papers and marking schemes and ends when the statistics of the finally corrected results are produced. The process comprises an inter-related set of procedures which altogether takes about two years to complete. The grade awarders who undertake their work near the end of the process have to assume that the question papers set at the beginning of the process present a fair test of the syllabus. They must also accept that the marking has been properly conducted so that each candidate's mark in each component is the best available measure of his or her performance in that component and thus that each candidate's aggregate mark (after any appropriate scaling has been completed) is the best available measure of his or her performance in the examination as a whole.

(b) Awarders are required to determine the minimum total mark required for each grade. Grading is essentially a comparative process with the aim of placing this year's grade boundaries at points on this year's mark scale which represent levels of performance equivalent to those required for the same grades in previous examinations (modified where appropriate in the light of any evidence of leniency or severity in previous examinations). Awarders are required to arrive at their decisions primarily through the study of scripts or other work submitted by candidates, and then by considering such other supplementary information as may be available.

(c) The three different sets of people present - SAC members, Chief Examiner(s)* and Subject Officer(s) - each have a particular kind of information to provide and when each has had an opportunity to do this a discussion of the evidence should lead to a consensus about the placement of the boundaries. In the event of minor uncertainty when no evidence to resolve the uncertainty can be found, a majority view is acceptable. In the event of major dispute, when all the resources of the Grade Awarding Panel have been exhausted and all the help and advice which can be gathered from Senior Staff and Research Officers has failed to resolve the difficulty, the ultimate responsibility for a decision has to rest with the SAC representatives.

(d) The task of the awarders is onerous. It involves careful preparation for the Grade Awarding Meeting itself, the exercise of expert individual judgement and the ability to participate properly in team work which will lead to a final collective judgement. As in most activities, the better the preparatory work is done, the easier the final task will be and the more smoothly the Grade Awarding Meeting will go.

Preparatory Work

The preparatory work described below is that which will normally take place in a subject where the examination is of a fairly conventional type, i.e. mainly written papers. In subjects where a high proportion of the marks is allocated for practical or oral components, or where, as for example in Art, the grading system has to be completely different in nature, corresponding activities should be undertaken wherever possible.

(a) <u>The Subject Officer</u>

The Subject Officer should prepare a table giving information about the entry for several past examinations, including numbers of candidates, distribution of types of candidates and types of centres, and any available information about changes in the centres presenting candidates.

* *The requirements of some syllabuses may demand the presence at Grade Awarding Meetings of other people with responsibilities closely related to those of the Chief Examiner. These could include the Co-Chief Examiners, Assistant Chief Examiners or Chief Moderators for coursework or projects. The Reviser or one or more Senior Assistant Examiners might be invited if the circumstances of the subject are such that they can make a unique and significant contribution. However, care should always be taken to ensure that the presence of such additional people does not unduly unbalance the Grade Awarding Panel.*

A further table should be prepared giving information and statistics about the same past examinations, raw mark allocations and any scaling factors for each component, component means and standard deviations, overall means and standard deviations, final grade boundaries (both component and overall if component grading is used) and percentages of candidates obtaining each grade.

The Subject Officer should collect together all the corresponding information and statistics for the current examination apart from the grade boundaries and grade statistics yet to be determined and should make a careful note of any peculiarities in these, e.g. where distributions do not include the full range of candidates because of late scripts or re-marks etc.

Finally, the Subject Officer should collect together all ancillary material relevant to the Grade Awarding Meeting - interboard comparability reports, School Examinations and Assessment Council Scrutiny Reports, SAC instructions, Education Committee instructions or comments, reports of any internal investigations on standards or other relevant studies, and any other 'official' documentation.

The Subject Officer must provide the materials which will enable the SAC members to conduct their preparatory work - past papers, marking schemes, archive scripts and scripts from the current examination, etc. [see (c) below]. In selecting scripts from the current examination the Subject Officer should take account of (i) the initial recommendations for grade boundaries made by the Chief Examiner [see (b) below]; (ii) the likely mark ranges for the various grades suggested by the statistics of previous years. A good supply of scripts should be provided from as wide a selection of centres as is possible within the Pre Mod Check List (or extended version of this if necessary).

(b) The Chief Examiner

The Chief Examiner should prepare initial ideas about possible grade boundaries either by considering his/her experience of the examination and its marking together with reports received from his/her Senior Assistant and/or Assistant Examiners or from a specially conducted scrutiny of scripts. A combination of both methods is likely to bring the most useful outcome. These initial ideas should be transmitted to the Subject Officer well before the Grade Awarding Meeting in order that they can be used as part of the evidence for selection of scripts for further preparatory work by the Chief Examiner and for the work of the SAC members before and at the meeting. The Chief Examiner should then undertake similar further preparatory work in the same way as the SAC members [see (c) below].

(c) The SAC Members

Prior to the meeting together of all awarders the SAC members should be provided with selected scripts (at grade boundaries A/B, B/C and E/N) from the previous year's examination together with the corresponding question papers, marking schemes and details of raw mark allocations, scaling factors applied and grade boundaries. They should be asked to study these in order to remind themselves of the standards applied in that examination. They should also be supplied with question papers, marking schemes, mark allocations and scaling factors for the current examination. In those subjects which have objective marking schemes it should be possible, by inspection, to estimate the robable grade boundaries. With criterion based marking schemes it should also be possible to estimate the boundaries from the descriptions of the mark bands for each question and, in subjects where the criteria for marking remain the same from year to year for a particular component, the grade boundaries for that component can be expected to remain the same. If at all possible, SAC members should also inspect selected scripts from the current examination before the meeting commences so that they enter into the discussions with some tentative notion of where this year's grade boundaries should be placed.

The Grade Awarding Meeting

The Grade Awarding Meeting should be attended by the Chairman of the Standing Advisory Committee (or his/her nominated alternate) together with two nominated members of the Standing Advisory Committee (a minimum of two members in addition to the Chairman is currently required), the Chief Examiner(s) for the subject and the Subject Officer. Members of the Senior Staff and Research Officers may be called in at any time to give advice or help with problems as may be required.

At the meeting it is necessary to determine by the full information/discussion process, the A/B, B/C and E/N boundaries. In each case the Subject Officer should then calculate the remaining boundaries by interpolation (see Appendix) and check whether the resultant pattern accords reasonably well with the pattern of previous years. If it does not then the Grade Awarding Panel should be asked to find a legitimate explanation of the changes. If it cannot, then it should be asked to consider amending the original grade boundaries. Interpolated grades must not be separately amended. The check against previous years can only be in respect of grades A, B and E.

In general there will be two kinds of activity at a Grade Awarding Meeting, (i) discussion of the preliminary notions of the awarders which they have brought to the meeting and of the conclusions which they have drawn from the evidence provided from various sources at the meeting, and (ii) the quiet review of the evidence, primarily of scripts from the current examination, conducted by each individual separately. The order in which these activities take place is not prescribed and will vary according to the nature of the examination and to the inclination of the members of the Panel. However, it is generally advisable to have an outline agenda for the meeting in order to ensure that all the necessary steps are taken. The first two items on the agenda should be standard. These are:

1. The Chairman should open the meeting having been briefed so that he/she can remind the members of the task which has to be undertaken, indicating where appropriate the changes to the 'A' Level grade scheme introduced in 1987. The Chairman may also wish to draw attention to the ancillary materials which the Subject Officer will have provided such as any comparability or scrutiny reports etc.

2. The Subject Officer should then remind the members about the details of the examination itself such as the mark allocations per component, any scaling factors which have been applied in order to achieve the weightings set out in the syllabus and should provide the members with any other information which it is appropriate to consider before the detailed work of the day begins.

The order for the rest of the business may vary but should include the following. If the grading is done by components then:

(a) The Chief Examiner responsible for the first component should report on the candidates' response to it and the members should then discuss this report in the light of their knowledge of the current and past papers and their particular views about possible grade boundaries formed during their preparations for the meeting.

(b) The awarders should come to preliminary conclusions about grade boundaries for the current examination and should then check these conclusions by reference to scripts covering a range of marks on either side. By discussion and by calling for additional scripts if necessary they should reach agreement about the grade boundaries. In practice it is likely that awarders will agree that all scripts above a certain mark are worthy of a particular grade while all scripts below some other mark are not and that the provisional grade boundary for that grade is the mark mid-way between these two marks.

(c) This procedure is then repeated for each of the other specified grades for the first component.

(d) The same procedure is repeated for all other components for which scripts or other material evidence is available. The Subject Officer should provide detailed guidance on the interpretation of any other material evidence, for example, that for objective tests, coursework, projects, orals, etc.

(e) Where no material evidence is available it will be necessary to receive a report from the Chief Examiner responsible for the particular component and this report should describe examples of the standard of work achieved by candidates. It is possible that other awarders may be able to contribute information relevant to judgements about this particular component and these must also be fully taken into account.

(f) The several component grade boundaries are added together having first been scaled as may be necessary to achieve the required weightings of the components. The awarders must then scrutinise the total examination work of a number of candidates who have achieved minimum overall marks for each of the specified grades in order to satisfy themselves that the overall standard is correct. At this stage it is worthwhile putting forward some of the statistical information about current and past examination. The component means and standard deviations may provide additional information about the way candidates have responded to the examination. Details of the component grade boundaries over recent years might also be supplied. (The percentage of candidates achieving each grade in the previous year's examination should NOT be revealed at this stage.) Subject Officers should draw the awarders' attention to anomalies between this year's provisional grade boundaries and those of previous years. It would be difficult to justify placing, say, the A/B grade boundary at a higher mark than was chosen last year whilst, at the same time, placing the E/N grade boundary at a lower mark than last year. Again, it would be noteworthy if the mean for one component had increased whilst that for another had decreased, when compared with the previous year. Another piece of statistical information that should be utilised is the cumulative percentages of candidates at each of the component grade boundaries for the current examination. These percentages should not differ too markedly between components unless there is evidence that the candidate entry as a whole has performed, in absolute terms, unusually well or unusually badly in a particular component. Incidentally, it must be recognised that regression-to-the-mean effects, arising from imperfect inter-paper correlations will result in cumulative percentages of candidates at the overall grade boundaries which are, in general, different from those at the component grade boundaries.

At 'A' and 'AS' Level all comparisons, whether of component marks or total marks, must be restricted to grades A, B and E only.

(g) After all the statistical evidence such as that mentioned above, including statistical indicators of possible changes in entry pattern, has been considered, the subject distribution should be consulted to see how the percentages of candidates in the various grades compare with last year's percentages. If there are marked differences which cannot be explained, then the awarders should be asked to repeat the process of establishing grade boundaries and to check previous standards by further scrutiny of the archive scripts.

The purpose of these further checks is not to persuade the awarders to change their decisions but to re-check them and only to change them if, on reflection, they feel that they were wrong. Once the specified boundaries have been established the remaining boundaries can be obtained by interpolation/extrapolation by the Subject Officer with the assistance of the general guidelines (see Appendix). These interpolations should be made before the Grade Awarding Panel disperses in order that checks and further work if necessary (see above) can be carried out.

In those subjects where holistic grading is used the preliminary work undertaken by the SAC members should include the scrutiny of the whole work of candidates as represented in the archive scripts. At the meeting itself there will be a discussion about the grade boundaries for the examination as a whole conducted along much the same lines as that for each individual component in the component grading system. The Subject Officer will probably have to prepare a supply of scripts covering a wider range of marks on either side of each of the specified boundaries and should attempt to choose scripts from candidates who have performed in a reasonably consistent way in all the components of the examination. If it is not possible to find real candidates who have achieved such comparability then notional candidates can be used taking work from different candidates which together gives the required total mark. If in a subject where the grading is conducted holistically the different components behave in widely varying ways such as one component having a very high mean mark while another has a very low mean mark then it may be necessary to take special steps to accommodate this. Certainly if it is an unusual occurrence in the subject then the advice of the Research and Statistics Division should be sought.

Indeed, if any problem arises during the course of any Grade Awarding Meeting then, according to the nature of the problem, a member of the Senior Staff or a Research Officer should be called in. If the grading which is proposed differs significantly from gradings in previous years then, again, a member of the Senior Staff must be called in before final confirmation is agreed.

Finally, at the end of the meeting, the decisions of the awarders should be recorded clearly on paper and a check made with the members that they are all agreed that this written record is correct. Immediately after the meeting, the Subject Officer will complete the grading scheme with information about grades A, B and E, for submission to Data Section and thus the production of computer printout giving the full grading scheme. All grade boundaries, including those interpolated or extrapolated, should then be checked. The printout should then be returned to Data Section, in order that the Moderation Check List may be produced.

GCE ADVANCED LEVEL

GENERAL GUIDELINES ON THE JUDGEMENTAL AND ARITHMETICAL FIXING OF GRADE BOUNDARIES

The Department of Education and Science has decided that, with effect from the Summer 1987 'A' Level examinations, grading standards will be established in the manner shown below.

(i) The A/B grade borderline should be established on the basis of Boards' existing practice, including examiners' judgement of quality.

(ii) The B/C and E/N grade borderlines should be established by reference to examiners' judgements of quality and using methods to ensure fairness and reliability at these crucial points.

(iii) The mark range between the B/C and E/N grade borderlines should be divided into three equal intervals and these assigned respectively to grades C, D and E, with the same interval also assigned to grade N.

How this will be done is described below.

(a) The A/B, B/C and E/N grade boundaries will be fixed on the basis of grade awarders' judgements.

(b) The C/D and D/E grade boundaries will be fixed by equipartitioning the mark range between the lowest mark for grade B and lowest mark for grade E. (If the number of marks is not exactly divisible by three, the following will apply:

- where there is a remainder of one, after the B/C to E/N mark range is divided by three, then the one extra mark is added to the grade C range;
- where there is a remainder of two, one extra mark is added to both the grade C and grade D ranges. Examples of these cases are given below.)

(c) The N/U grade boundary will be fixed by extrapolating from the E/N grade boundary so that the mark range for grade N is the same as that for grade E.

Example 1

Consider an examination with a maximum mark allocation of 100 marks. The grade boundaries would be arrived at as described below.

1. To maintain the standard set in previous years, the awarders have decided that 70 should be the lowest mark for grade A.

2. To maintain the standard set in previous years, the awarders have decided that 60 should be the lowest mark for grade B.

3. To maintain the standard set in previous years, the awarders have decided that 41 should be the lowest mark for grade E.

4. The mark range between the B/C and E/N grade boundaries is 60 - 41 = 19 marks.

5. Since the mark ranges for grades C, D and E have to be equal, the C/D and D/E grade boundaries are arrived at in this case as follows.

   (i) The mark range for each of these three grades is $\frac{19}{3}$ = 6, remainder 1.

   (ii) Now, to give the candidates the maximum possible benefit, the extra mark is added to the grade C range, giving mark ranges of 7, 6 and 6 for grades C, D and E respectively.

   (iii) The lowest mark for grade C is 60 - 7 = 53 and for grade D is 53 - 6 = 47.

6. Since the mark range for grade N has to be the same as that for grade E, the lowest mark for grade N is obtained by subtracting 6 (the mark range for grade E) from 41, the lowest mark for grade E, giving 35.

The results are summarised in tabular form below.

Grade	Lowest Mark	Mark Range	Method
A	70	100 - 70	Inspection
B	60	69 - 60	Inspection
C	53	59 - 53	Interpolation
D	47	52 - 47	Interpolation
E	41	46 - 41	Inspection
N	35	40 - 35	Extrapolation
U	0	34 - 0	-

Example 2

Consider an examination with a maximum mark allocation of 100 marks. The grade boundaries would be arrived at as described below.

1. To maintain the standard set in previous years, the awarders have decided that 69 should be the lowest mark for grade A.

2. To maintain the standard set in previous years, the awarders have decided that 61 should be the lowest mark for grade B.

3. To maintain the standard set in previous years, the awarders have decided that 41 should be the lowest mark for grade E.

4. The mark range between the B/C and E/N grade boundaries is 61 - 41 = 20 marks.

5. Since the mark ranges for grades C, D and E have to be equal, the C/D and D/E grade boundaries are arrived at in this case as follows.

   (i) The mark range for each of these three grades is $\frac{20}{3}$ = 6, remainder 2.

   (ii) Now, to give the candidates the maximum possible benefit, the extra marks are added to the grade C and D ranges, giving mark ranges of 7, 7 and 6 for grades C, D and E respectively.

   (iii) The lowest mark for grade C is 61 - 7 = 54 and for grade D is 54 - 7 = 47.

6. Since the mark range for grade N has to be the same as that for grade E, the lowest mark for grade N is obtained by subtracting 6 (the mark range for grade E) from 41, the lowest mark for grade E, giving 35.

The results are summarised in tabular form below.

Grade	Lowest Mark	Mark Range	Method
A	69	100 - 69	Inspection
B	61	68 - 61	Inspection
C	54	60 - 54	Interpolation
D	47	53 - 47	Interpolation
E	41	46 - 41	Inspection
N	35	40 - 35	Extrapolation
U	0	34 - 0	-

APPENDIX 5.2

AWARDING PROCEDURE DOCUMENT IN USE DURING PHASE 3

REVISED JUNE 1993
Procedure File No 8

AWARDING

GCE ADVANCED LEVEL AND ADVANCED SUPPLEMENTARY

EXAMINATIONS IN 1993

CONTENTS

a:pr6con

Procedure File No. 8

AWARDING
GCE ADVANCED LEVEL AND ADVANCED SUPPLEMENTARY EXAMINATIONS IN 1993

1. INTRODUCTION

This document prescribes the procedures for determining grade boundaries for Advanced Level and AS examinations in 1993. It does not include such matters as possible training for grade awarders, aegrotat awards, borderline re-marking, cheating cases or other matters to do with assigning or issuing grades. These are covered in other documents.

The determination of grade boundaries is a crucial part of the examination process which begins with the setting of question papers and marking schemes and ends when the statistics of the finally correct results are produced. The process as a whole comprises an inter-related set of procedures which altogether takes about two years to complete. The grade awarders who undertake their work near the end of the process have to assume that the question papers set at the beginning of the process present a fair test of the syllabus. They must also accept that the marking has been properly conducted so that each candidate's mark in each component is the best available measure of his or her performance in that component and thus that each candidate's aggregate mark (after any appropriate scaling has been completed) is the best available measure of his or her performance in the examination as a whole.

Awarders are required to recommend the minimum mark required for each of three Key grades: A, B and E, for each component of the examination separately. Grading is essentially a comparative procedure with the aim of placing this year's grade boundaries at points on this year's mark scales which represent levels of performance equivalent to those required for the same grades in previous examinations (modified where appropriate in the light of any evidence of leniency or severity in previous examinations). Awarders are required to arrive at their recommendations on the basis of two forms of evidence:

a. their professional judgements of how the quality of the work found in scripts or other work submitted by the current candidates compares with the quality of the work of previous candidates;

b. the statistical data which are available showing how the marks awarded in the current examination compare with those awarded in previous years.

2. PREPARATORY WORK

The preparatory work described below is that which is required in a subject where the examination is of a fairly conventional type, i.e. mainly written papers. In subjects where a high proportion of the marks is allocated for practical or oral components or where, as for example in Art, the grading procedure has to be completely different in nature, corresponding activities must be undertaken.

In preparation for the awarding meeting, the Subject Officer should carry out the following tasks, consulting as necessary with the Research Officer or senior member of Education Division staff assigned to support the meeting in question (*the support officer*).

a. Statistical Reference Year

Identify which of the preceding three years' examinations is to be used as a reference for the current year's award. This must be the immediately preceding year's examination unless there is good reason to believe that the grade standards set in the preceding year were inappropriate. Advice should be sought from the support officer or from another senior member of staff of Education Division in this case. A written rationale must be prepared by the Subject Officer explaining the decision if any other than the preceding year's examination is used as the statistical reference. This rationale must be filed with the other papers from the awarding meeting so as to inform any subsequent investigation of the standards used in the year in question.

b. Agenda

Prepare an agenda for the meeting (an outline agenda is attached as Appendix 1).

c. Despatch to all those attending the Meeting

Despatch to all persons (including the support officer) who will attend the meeting, the agenda and the following materials and background information to enable them to prepare for it:

(i) question papers, marking schemes and archive scripts illustrating the standards which were set for each of the key grade boundaries on each component in the previous year's (or other reference) examination;

(ii) details of the component raw mark allocations, mean marks, standard deviations and scaling factors for the previous year's (or other reference) examination;

(iii) the component and subject boundaries used to award the key grades in the previous year s (or other reference) examination, together with the cumulative percentages of candidates at each of them. (Note that, if the recommendations of the awarding committee in the reference year were not accepted, it still is the **boundaries actually used to award grades in the reference year** which are sent out to the committee before the current year s meeting.);

(iv) question papers and marking schemes from the current examination;

(v) details of the component raw mark allocations and scaling factors for the current examination;

(vi) the awarders' briefing paper (attached as Appendix 2).

d. Statistically equivalent boundaries (SEBs)

There are two different approaches to finding the marks on a particular component in the current year's examination which are equivalent to the grade boundaries adopted for that component in the reference year. Which of these approaches is used depends upon the nature of the component in question.

(i) Components which are used only in a single normal enterable A-Level subject

(Note: this category **also includes** components which are common only to
an A-Level and an A-Level + S paper combination;
an A-Level and AS examination;
or a normal A-Level and its special overseas version.

In any of these cases, the following approach should be used to determine Statistically Equivalent Boundaries, based only on the data for the normal A-Level examination.)*

Compare the unit mark distribution for the current main A-Level examination with the corresponding one from the previous year's examination. (Unless it has been decided to use another of the preceding three years' examinations as a reference, in which case the comparison should be with the chosen examination.) Check that the data for the current examination are comparably complete with those from the previous year. (If there are doubts about the comparability of the data from the two years advice should be sought from the Research and Statistics Group.) Identify the three marks which give the same cumulative proportions of candidates as were obtained on the component the year before at the three key grade boundaries used to award grades (i.e. not those recommended by the awarding meeting, if these were different). These marks are the statistical equivalents, for the current year, of last year's key grade boundaries for the component in question.

(ii) Components which are used in more than one normal enterable A-Level subject

(Note: The components covered by this category in 1993 are listed in Appendix 7.
This category does not include components which are common only to
an A-Level and an A-Level + S paper combination;
an A-Level and AS examination;
or a normal A-Level and its special overseas version.

* but see Appendix 8 for examinations with special overseas versions

In these cases, the approach described above in Section d(i) should be used.)

For components which are used in more than one normal enterable A-Level examination (see Appendix 7), two or more unit distributions will be available since the component will appear as a unit in two or more subjects. In these circumstances, the process described above, if repeated for each unit distribution relating to a particular component, will produce two or more estimates for each SEB. These estimates are usually the same or close to each other. However, in subjects with several options, some taken by only a small number of candidates, this may not be so. It is therefore better to base the Statistically Equivalent Boundaries upon component, rather than unit, distributions.

A Subject Officer with a component listed in Appendix 8 should therefore request from Computing Services Division a component distribution for that common component in addition to the unit distributions needed for the awarding meeting. The component distribution is then used in place of the unit distribution in the procedure described above in Section d(i) to identify the statistically equivalent boundaries for the component. Since this procedure is new for 1993, corresponding 1992 component distributions are not available among last year's awarding data. These will, however, be supplied by Computing Services Division for all the components listed in Appendix 8.

e. Starting ranges

For each component, discuss initial suggestions for the location of the key grade boundaries (A/B, B/C and E/N) with the Chief Examiner (or the Senior Examiner responsible for the component). The Chief Examiner will have initial ideas about possible grade boundaries based upon his/her experience of the examination and its marking, together with reports received from his/her Senior Examiner and/or Assistant Examiners or, perhaps, from a specially conducted scrutiny of scripts. Discuss these initial ideas in the light of the statistically equivalent boundaries identified in (d) above.

By this means, arrive at three agreed starting ranges of marks (for each component) within which the key grade boundaries are likely to lie for the current examination. Each range must include marks either side of the statistically equivalent boundaries and should span at least 10% of the mark range for the component or at least 10 marks if the component is marked out of more than 100 marks. If possible, the starting ranges should also include the Chief Examiner's initial suggestions. However, the Subject Officer has the final decision about the starting ranges to be used at each boundary.

f. Scripts for the Meeting

Well before the meeting, arrange with Scripts Department for the extraction and delivery of large samples of scripts for each component. A good supply of scripts should be provided from as wide a selection of centres as is possible within the Range Mark-Grade List (or full

version of this if necessary). Once the starting ranges have been decided, scripts covering these ranges should be selected from the sample supplied by Scripts Department. Wherever possible, the selected scripts should show a balanced performance across the range of questions to be answered. At least two scripts on each mark within each starting range should be chosen unless there are no such scripts available. Scripts for which an examiner adjustment has been applied should not be included in the sample used for awarding purposes. For components other than written papers, several samples of candidates' work near each of the statistically equivalent boundaries (e.g. tapes of orals or project folders) should be available wherever possible.

g. Statistical Information for the Meeting

Gather together, for use at the meeting, the following statistical information in addition to that sent to those attending the meeting (see c. above):

(i) the unit and subject mark distributions and the mark-grade list for the current year's examination;

(ii) the unit and subject mark distributions used in the awarding meeting of the previous year's examination (or the reference examination where this is not the immediately preceding one)*;

(iii) historical data for the subject concerned in the previous five years (these should include the distributions of different types of candidates and centres and the final percentages of candidates awarded each grade - suitable sources are the Subject Statistics which are received from the Research and Statistics Group in December of each year and include other useful information such as the means and standard deviations of the marks or, since Subject Statistics are not available for examinations held before 1989, the Board's published Statistics).

h. Awarding Pro-forma

Arrange for a supply of multiple copies of the standard awarding pro-forma upon which the awarders record their judgement of each script which they scrutinise (copy attached as Appendix 3).

i. Ancillary Material

Gather together all ancillary material relevant to the Grade Awarding Meeting - interboard comparability reports, School Examinations and Assessment Council (SEAC) Scrutiny Reports, SAC instructions, Education Committee instructions or comments, reports of any internal investigations on standards or other relevant studies, and any other similar documentation.

* but see Appendix 8 for examinations with special overseas versions

3. THE AWARDING MEETING

This section describes the roles of the various personnel involved in grade awarding meetings, the outcomes they are required to produce and the procedures to be used in 1993.

3.1 Personnel and Responsibilities

The awarding meeting should be attended by:

a. the Chairman of the Standing Advisory Committee (or his/her nominated alternate) together with at least two other members of the Standing Advisory Committee (i.e. a minimum of three SAC members including the Chairman of the awarding meeting is required);

b. the Chief Examiner(s) for the subject (including, as appropriate, Senior Examiners who have primary responsibility for any component, Senior Coursework Moderators and so on);

c. the Subject Officer.

The designated support officer and other members of the Senior Staff of Education Division will also attend the meeting to give any help or guidance, which they judge necessary. Such staff may also be called into the meeting at any time if a problem arises.

The three different sets of people required to attend awarding meetings - SAC members, Senior Examiners and the Subject Officer - each have a particular kind of expertise and information to contribute. Discussion of this evidence should lead to recommendations about the placement of the key grade boundaries which represent the consensus of the meeting. In the event of minor disagreement when no evidence to resolve it can be found, a majority recommendation is acceptable. In the event of major dispute, when all the resources of the Grade Awarding Meeting have been exhausted and all the help and advice which can be gathered from Senior Staff and Research Officers has failed to resolve the difficulty, the responsibility for arriving at final recommendations rests with the SAC representatives.

The awarding meeting makes its recommendations to the Board. Provided that these recommendations are consistent with the Board's established expectations (see Section 3.5 Paragraph e., below) they will normally be accepted unchanged. On occasions when the recommended grade boundaries are not consistent with established expectations, the Board, which is represented for this purpose by the Secretary General, will consider the recommendations, the reasons for them and any other relevant matters before deciding upon the boundaries which should be used to award candidates' grades in the examination in question (see Section 3.8, below).

3.2 The Outcomes of the Meeting

The main outcomes of the grade awarding meeting are the recommendations for the key grade boundaries (A/B, B/C and E/N) for each component. Once approved, these are combined, by computer, in a prescribed way (described in Section 3.5) to give overall Subject boundaries for the examination as a whole. Should the recommended component boundaries imply changes in the cumulative percentages of candidates at the key Subject grade boundaries which are greater than the expected maxima established by the Board, then they must be accompanied by a written rationale **prepared before the awarding meeting disperses.** (The written rationale is discussed in more detail in Section 3.6.)

3.3 Initial Procedures

a. Chairman s Introduction

The Chairman should open the meeting by reminding the members of the task which has to be undertaken and outlining the procedures to be followed. The Chairman may also wish to draw attention to the ancillary materials which the Subject Officer may have provided such as any comparability or scrutiny reports.

b. Subject Officer's Introduction

The Subject Officer should then deal with domestic matters and remind the members about the details of the procedures and of the examination itself. These will include the mark allocations for each component and any scaling factors which have been applied in order to achieve the weightings set out in the syllabus and should provide the members with any other information about the examination as a whole which it is appropriate to consider before the detailed work of the day begins. The Subject Officer will also give the meeting details about the characteristics of the entry for the current year and the previous reference year.

3.4 Establishing provisional key boundaries for the components

a. Order of components

The order in which the components are tackled is not prescribed except that externally marked written papers, where they exist, must be considered before all other types of component.

b. Senior Examiner s report

The Senior Examiner responsible for the first component begins its consideration by reporting on the way in which it has functioned. In particular, he or she should refer to any questions which were easier or harder for candidates to respond to than expected. If there are any questions in the current examination for which there were closely

related counterparts in the reference year's examination, then the Senior Examiner should point out any perceived differences in the questions themselves which would have affected the ease with which candidates could score marks on such questions compared with their previous counterparts. If the marking scheme used in the current year differs from the reference year in the proportion of marks awarded to a particular skill, set of skills or class of knowledge, then the probable effects of this upon the marks awarded should be explained. The essence of the Senior Examiner's report should not be his or her judgement of how the candidates have performed but, rather, the effect of specific features of particular questions (which either presented particular difficulties to the candidates or which encouraged particularly good responses), together with the current marking scheme, upon the ease with which candidates gained marks in the current examination.

c. Subject Officer's report

The Subject Officer then outlines the statistical evidence concerning the difficulty of obtaining marks in the first component in this year's examination compared with the same component in the previous year's examination (or the reference examination being used if this is not the immediately preceding one). The Subject Officer should compare the means and standard deviations of the marks for the component in the two years and draw out the implications for the difficulty of the component in the current year's examination.

d. The first boundary: E/N

The meeting then turns to the determination of the first component grade boundary: E/N. The Subject Officer informs the meeting of the mark which is statistically equivalent to the lowest mark in the previous year's Grade E for this component (see Section 2, Paragraph d.) and of the starting range of marks which has therefore been chosen as likely to contain the E/N boundary on this component (Section 2, Paragraph e.).

e. Scrutiny of Scripts

The members of the meeting now begin scrutinising scripts within the starting range with the purpose of identifying, in their professional judgement, which mark is attached to scripts which represent the same standard as those with the lowest component mark in Grade E in the previous year. Where sufficient scripts are available, it is helpful if each script is scrutinized by no more than two of the awarders. If this can be arranged, it ensures that a reasonable number of scripts are considered by the meeting as a whole and inhibits excessive discussion of individual, possibly unrepresentative, examples of candidates' work. For each script which they scrutinise, the awarders record the grade which they believe it to be worth on the pro-

forms provided for this purpose (see Appendix 3). The essence of the process is comparative; the task is to judge this year's scripts, as responses to this year's question paper marked according to this year's marking scheme, using standards inferred from last year's archive scripts, question paper and marking schemes for the same component. This is the case even for a new or revised syllabus although the comparison is then more difficult to make, being with a different syllabus in the same subject area rather than with the same component on a different occasion.

f. The initial provisional recommendation

Once each member is satisfied that he or she has reached a tentative conclusion about the location of the grade boundary, the Chairman calls the meeting to order. There then follows a discussion of the results of the script scrutiny and of the statistical evidence presented earlier by the Subject Officer. This discussion may lead to the adoption of a provisional recommendation for the location of the E/N boundary for the component. It is likely, however, that awarders will not all be able to agree on a single mark, but that they will be able to agree that all scripts above a certain mark are worthy of Grade E and that all scripts below some other mark are not. The establishment of such a borderline mark-range is all that is required at this stage in the procedure.

g. The other key grade boundaries: B/C and A/B

The procedures specified in d. to f. above are now repeated for the other two key boundaries B/C and A/B, in that order.

h. Review of provisional recommendations for the component

Once provisional recommendations for all three key boundaries (or provisional borderline mark-ranges) have been agreed, the results for the component as a whole should be reviewed. At this stage, useful checks upon the consistency of the qualitative judgements at the three grade boundaries can be made by the Subject Officer and discussed in the meeting. For example, comparing the details of the component grade boundaries with the previous year may show up anomalies between this year's provisional grade boundaries and the confirmed grade boundaries of last year; it would be difficult to justify placing the A/B grade boundary for a component at a higher mark than was chosen last year whilst, at the same time, placing the E/N grade boundary at a lower mark than last year unless the standard deviation of the marks had increased. Furthermore, as with subject boundaries (see Section 3.5, Paragraph e.) in examinations with entries of over 500 candidates, the cumulative percentages of candidates at the component grade boundaries for the current examination should not differ by more than a few percentage points from those in the previous year unless there is some independent reason to believe that the candidates, as a group, are more or less competent than last year's. Certainly, a large change in

the cumulative percentage of candidates above, say, the B/C boundary which was unaccompanied by a corresponding change in the percentage above the A/B boundary would be suspect. Considerations of this kind may lead to further modification and refinement of the provisionally agreed component grade boundaries. They also help an appropriate value to be chosen in those cases where a borderline mark-range, rather than a single boundary mark, was agreed after the scrutiny of scripts.

i. The other components

The procedures set out in Paragraphs b. to h. above are then repeated for all other components for which scripts or other material evidence is available. The Subject Officer should provide detailed guidance on the interpretation of work produced in response to components other than externally marked written papers (for example, coursework, projects or orals).

(i) Coursework

For coursework components, where the work of each candidate is characteristically fairly extensive, it is not normally possible for the awarders to scrutinise several examples of work on each mark point and the conclusions which they reach from the scrutiny of candidates' work are therefore less reliable than is the case with externally marked written components. For coursework components, therefore, greater emphasis should be placed upon the statistical evidence than in the case of externally marked written papers. Where the requirements for a coursework component have been specified and the marks ascribed on the same basis as hitherto, there is no reason why the grade boundaries should alter much from year to year. Where requirements or marking criteria have changed, however, the statistically equivalent boundaries should normally be used.

(ii) Components where the work is ephemeral

For components where no candidates' work is available for scrutiny at all, it is necessary to receive a report from the Senior Examiner responsible for the particular component describing examples of the standard of work achieved by candidates. In such cases, the principal evidence which should be used to fix the grade boundaries is the statistical evidence. As with coursework, where the requirements for such a component have been specified and the marks ascribed on the same basis as hitherto, there is no reason why the grade boundaries should alter much from the previous year. Where requirements or marking criteria have changed, however, the statistically equivalent boundaries should normally be used.

(iii) Objective Tests

The case of Objective Test components is a special one. Setting comparable grade boundaries on such tests by other than purely statistical means is generally accepted to be extremely difficult. It is not, therefore, part of the Board's procedures for the awarding meeting to set boundaries for Objective Test components. In examinations which use them, the process of determining subject boundaries is modified in such a way as to render judgements about Objective Test component boundaries unnecessary (see Appendix 4). Nevertheless, since the candidates' individual grades for the subject as a whole are arrived at by aggregating the marks for all components (including any Objective Test), performance in the Objective Test contributes in the normal way to each candidate's subject total and hence to the grade which they achieve.

(iv) Special overseas papers

The grading procedures for these papers are set out in Appendix 8. The awarding meeting is not involved because the grading is done entirely statistically on the basis of the grade boundaries set by the awarders on the corresponding papers for home candidates.

3.5 Checking the Subject Boundaries

Once the three key boundaries have been provisionally determined for each component, it is necessary to check upon their effect for the subject as a whole. In particular, a check must be made that any changes from the previous year in the cumulative percentages of candidates at each of the key Subject grade boundaries do not exceed the Board's maximum expectations. (Certain special conditions apply in the case of subjects involving options; see Appendix 7.) To make this check, the provisional component boundaries must be combined to produce boundaries for the subject as a whole. There are two methods for doing this; both are used at each boundary and the lower one (which is exceeded by the marks of the greater number of candidates) is chosen.

a. <u>The percentile method</u>

For the E/N boundary, the Subject Officer first determines, from each unit distribution, the cumulative percentage of candidates whose marks exceed the provisional boundary mark on each component of the examination. The average (mean) of these percentages is then calculated (taking appropriate account of the weight given to each component in the syllabus) and the total mark for the subject as a whole which, as nearly as possible, is exceeded by this mean percentage of candidates is found.

This total mark is one possible E/N boundarymark for the subject as a whole. (Full computational details of this method are given in Appendix 4.)

b. The addition method

In this method, the Subject Officer simply adds together the provisional component boundaries using the normal scaling factors for the components of the examination. The result is a second possible Subject grade boundary.

c. Comparison

The Subject Officer then determines which of the two possible Subject E/N boundaries is the lower and this becomes the putative Subject E/N boundary.

d. The other key boundaries: B/C and A/B

The Subject Officer then repeats the procedures in Paragraphs a. to c. for the other two key boundaries B/C and A/B.

e. Comparison of statistical outcomes with the Board's expectations

For examinations attracting more than 500 candidates, the Board has established the maximum change in the cumulative percentage of candidates at each key Subject boundary which is to be expected in any one year. These are as shown below:

Grade	Expected Maximum annual change (cumulative percentage points)
A	1%
B	2%
E	3%

The Subject Officer now checks the cumulative percentage changes from the previous year's results which follow from the putative Subject boundaries determined in Paragraph c. above. (If the reference examination is not from the immediately preceding year, the comparison is with the year in which it was taken.) The purpose of this check is to see if the changes are within the expected maxima set out above. The action to be taken depending upon the result is prescribed in Paragraphs g. to i. below.

In 1993, the comparison of statistical outcomes is made more complex by the special phasing-in procedure used in 1992 for boundaries derived by the percentile method. For such boundaries (normally A/B and B/C), the cumulative percentage of candidates awarded the grade by the putative boundary should not be compared with the cumulative percentage actually awarded the grade in 1992. Rather, it must be compared with the cumulative percentage **which would have been awarded the grade if the phasing-in procedure had not been used for the boundary in question**

in 1992. (These cumulative percentages were computed as part of the 1992 awarding procedure and should, therefore, be available from the records of the 1992 awarding meetings. If they are not available for any reason they should be reconstructed from the 1992 data before the grading meeting begins. The Research and Statistics Group can offer help in this respect, if required.)

f. Small entry subjects

For examinations with fewer than 500 candidates, greater changes than those shown in Paragraph e. are likely to occur from year to year with greater frequency. As a result no limits are prescribed. However, the expected changes for examinations in this category are set out in Appendix 5. Awarding committees should bear these in mind when reviewing the statistical consequences of their recommendations; further action is not prescribed for small entry subjects but the general approach specified in Paragraphs g. to i. below should be followed as appropriate.

For small entry subjects covered by this section, the awarders' recommendations will normally be accepted on behalf of the Board by the Subject Officer, after consultation with the designated support officer for the meeting in question. However, either of these officers may refer the recommendations to the Secretary General (or his nominated alternate) if they judge it necessary.

g. If the cumulative percentage change is within the expected annual maximum at any grade

At any grade for which the cumulative percentage change for the Subject as a whole, based on the putative Subject grade boundary, is within the expected annual maximum, the provisional component boundaries can be adopted as final recommendations without further consideration and can be accepted on behalf of the Board by the Subject Officer.

h. If the cumulative percentage of candidates at a grade changes by more than the expected annual change

It is occasionally the case that a change in cumulative percentage exceeds the Board's established expectations for a statistical or technical reason. For example, in certain circumstances, the change in cumulative percentage for the subject as a whole can just exceed the established expectation even though the awarders recommend the statistically equivalent boundary for each component. Another example occurs in the case of subjects with common components (see Appendix 7) where boundaries which meet the expectations simultaneously for each such subject cannot be found. In these circumstances, advice must be sought from the designated support officer for the meeting who, if the explanation is technical, will assist the Subject Officer in drawing up a written statement explaining the reasons for exceeding the Board's established

expectations. Once this has been done, the provisional component boundaries can be adopted as firm recommendations and accepted on behalf of the Board by the Subject Officer. However, if they judge it necessary, either the Subject Officer or the support officer may refer the recommendations and technical explanation to the Secretary General (or his nominated alternate) for approval.

In the majority of cases, where there is no technical explanation for a cumulative percentage change greater than the expected annual maximum, the provisional component boundaries must be reconsidered by repeating Procedures e., f. and h. from Section 3.4. It may be the case that the provisional boundary at the grade in question for one of the components, in particular, is doubtful because there was considerable divergence between the awarders' professional judgements of the sample scripts or because the provisional boundary was fixed an unusually large number of marks away from the statistically equivalent boundary established by the Subject Officer (see Section 2, Paragraph d.). In cases like this, it makes sense to begin the reconsideration with the particularly doubtful component. Otherwise, each component must be reconsidered at the boundary in question. If the reconsidered component boundaries lead to a putative Subject boundary which gives a change in cumulative percentage within the expected maximum, then they can be adopted as firm recommendations and accepted on behalf of the Board by the Subject Officer.

i. If the cumulative percentage change is greater than the expected annual change after the provisional component boundaries have been reconsidered

If, after reconsideration of the provisional component boundaries, the cumulative percentage change is still greater than the expected annual change, then the awarders must discuss the situation with the designated support officer for the meeting and/or with a member of the Senior Staff of Education Division. If, after that discussion, the awarders still wish to maintain their decisions, the reconsidered component boundaries go forward from the meeting as the recommended boundaries. In this case, a written rationale for the recommended component boundaries must be prepared **before the meeting disperses.**

3.6 The Written Rationale

If an awarding panel carries out a reconsideration of their provisional component boundaries but still wishes to make recommendations which produce greater than expected changes in the cumulative percentage of candidates at one or more grades, then they may do so. However, they are required, in these circumstances, to accompany those recommendations with a written rationale which is **prepared before the meeting disperses.** The written rationale will normally be prepared by the Subject Officer, acting as Secretary to the awarding meeting. The rationale should be signed by the Chairman to indicate that it represents the

consensus view of the meeting. It is necessary for the written rationale to be detailed and specific. Bland *post hoc* justifications such as "the quality of the candidates' work is higher/lower this year" are not sufficient.

The written rationale must first address the issue of the relative difficulty of scoring marks in the current and reference year's examinations, supporting the view taken on this matter by explicit reference to the questions and marking schemes from the two examinations. It must also discuss the implications of the mark statistics for the relative difficulty of the current examination. The rationale should then turn to the performance of candidates at the recommended component boundaries and discuss in detail how this compares between the two years in terms of the features which are relevant to judgements of the quality of the work. From the scrutiny of candidates' scripts, there are likely to be insights about the nature of the work of the current year's candidates to which reference should be made. For example, in a physical science subject a deterioration in the capacity of candidates to understand and use algebraic expressions might be observed. Provided that the algebraic demands of the current papers were judged no greater than those of last year, then an effect of this kind could be quoted to support the view that an accompanying drop in the mean mark was the result of reduced performance by the candidates rather than a harder examination. As a result, lower percentages of candidates awarded the top grades could be expected. Other comments may also be made; for example, that one year's candidates are better/worse prepared than the other's. However, any such assertion **must be supported by evidence** in the form of relevant specific examples or a description of the relevant features of candidates' work from which this inference is drawn.

Preparing the rationale is likely to be time consuming and difficult and to be done at the end of a long, and possibly argumentative, day. However, recommendations for boundaries which produce greater than expected changes in the cumulative percentage of candidates at one or more grades must be accompanied by an appropriately detailed written rationale if they are to have any chance of being approved by the Board.

3.7 Interpolating the other Subject boundaries

Once the key subject boundaries have been established, the remaining subject boundaries can be obtained by interpolation/extrapolation by the Subject Officer following the national guidelines (see Appendix 6). It is not **essential** for these interpolations to be made at the meeting because they will eventually be made by computer. However, the awarders may ask the Subject Officer to do so in order to get as complete a picture of the results of their deliberations as possible.

3.8 Recording the recommendations and implementing the results of the Meeting

Finally, at the end of the meeting, the recommendations of the awarders should be recorded clearly on the standard form (a sample form is attached as Appendix 9) and a check made with the members that they are all agreed that this written record is correct. The Chairman and Subject Officer then both sign the form to indicate that this is so.

In cases where the Subject Officer has not been able to accept the meeting's recommendations on behalf of the Board (see Section 3.5 Paragraph i.) approval for them must be sought. **Immediately after the meeting,** the Subject Officer will therefore take the recommendations, together with the written rationale to the Secretary General for approval. (In the absence of the Secretary General, the Deputy Secretary General will be consulted and, in his absence, the Head of Education Division.) In such cases, the Secretary General (or his alternate) determines the grade boundaries which will be used to award grades to candidates. Each case is considered on its merits but, in the absence of a clear rationale, boundaries which give statistical outcomes as similar as possible to those of the reference year are most likely to be used.

The completion of the form recording the awarding meeting's recommendations marks the formal end of the awarding meeting. Awarders may, if they wish, remain to hear the Secretary General's decision but this decision is final. If the awarders are concerned about the grade boundaries finally adopted, these concerns should be raised when the responsible Standing Advisory Committee reviews the outcomes of the examination at its Autumn meeting. At that time, the Standing Advisory Committee can request an investigation of the standards of the awards to be carried out by the Research and Statistics Group. The results of any such investigation will be used to inform the following year's awarding meeting.

In cases where the Subject Officer has been able to accept the meeting's recommendations on behalf of the Board (see Section 3.5 Paragraphs f., g. and h.), he or she will **immediately** inform the Secretary General, his deputy or the Head of Education Division, as appropriate, that this is the case.

When an approved set of grade boundaries is available, they will be entered into the Board's computer and the process which selects candidates whose work is to be borderline re-marked will be initiated (see Procedure File No. 9).

3.9 Archiving of scripts for future awarding meetings

It is necessary to ensure that an adequate supply of archive scripts at the component boundaries is available for reference purposes at the following year's awarding meeting. It is sensible, in this respect, to identify scripts during the meeting which are suitable for this purpose and to make a note of their centre and candidate numbers for future extraction. On those rare occasions when the awarding meeting's recommended boundaries are not approved but are modified by the Board before implementation, it is important to ensure that archive scripts are retained at the boundary marks actually implemented and not at those originally recommended.

18th June 1993

PROFILES

APPENDIX 1

OUTLINE AWARDING MEETING AGENDA

1. Chairman s Introduction

2. Subject Officer's Introduction

3. Establishing of provisional key boundaries for the components

   First written paper

   E/N Senior Examiner's Report
   Subject Officer's Report
   Scripts Scrutiny and Initial Recommendation

   B/C

   A/B

   Review of initial component recommendations

   Other written papers

   Other types of component

4. Checking the Subject boundaries

5. Firming up the Component boundary recommendations

6. Recording the recommended boundaries

a:mcapp1

APPENDIX 2

BRIEFING PAPER FOR AWARDERS CONCERNED WITH A-LEVEL AND AS EXAMINATIONS IN 1993

1. INTRODUCTION

The determination of grade boundaries is the last major stage in the examination process prior to the issue of results. In some respects it can be regarded as the most important part of a cycle which began nearly two years previously, with the initial draft question papers devised by the Chief/Senior Examiner. The importance of the grade awarding process lies in the fact that the grades awarded constitute the end product which the centres and candidates receive. They affect the futures of the candidates and the services provided by the Board will be judged by the extent to which the results are seen to be fair and reliable.

Because grade awarding is such a critically important part of the Board's work, it is essential that the procedures for achieving it are valid, reliable and command public confidence. The purpose of this paper is to brief those involved in awarding A Level examinations in 1993 on the procedures to be used. First, however, Section 2 sets out the purposes of awarding meetings.

2. THE PURPOSES OF AWARDING MEETINGS

The purpose of any particular awarding meeting depends upon whether or not it is the first held for a particular syllabus. If the first examination on a new syllabus is being awarded, then the awarding meeting is required to establish the grading standards to be applied to that syllabus' examinations on that first occasion and thereafter. In doing this, it is important for due regard to be paid to the standards required for each grade on examinations of other syllabuses in the same subject and, indeed, at A-Level in general.

For awarding meetings other than the first, the purpose is to apply the grading standards established at the first awarding meeting to each succeeding year's examination. (In the case where there is evidence that previous standards require adjustment, the awarding meeting at which the adjustments are made has essentially the same purpose to the first awarding meeting for a syllabus.)

In both cases, the essence of the awarding meeting is the maintenance of comparable standards. Awarding meetings are the key device through which examining boards discharge their public duty to maintain standards. The maintenance of comparable standards between years within a syllabus and between different syllabuses in the same subject is one of the

necessary features of a fair examining system. This is because candidates who obtain their qualifications in different years or by following different syllabuses subsequently compete for the same jobs and places in Further and Higher Education. For a new or substantially revised syllabus the question which the first awarding meeting is intended to answer is thus:

> *What mark would candidates who just barely achieve Grade x on other syllabuses in this subject have achieved in this examination?*

Annual awarding meetings are required because it is not possible to achieve precise consistency between successive years' examinations (upon a particular syllabus) in terms of the number of marks which represent a given quality of work. Therefore, it is necessary to establish afresh, for each year's examination, exactly what mark is equivalent to each grade boundary mark in the previous year. In essence, the awarding process enables allowance to be made for the, hopefully small, changes which inevitably occur in the difficulty of examinations (or in the marking standards applied) from year to year. The question which awarding meetings for an established syllabus are intended to answer is thus:

> *What mark would last year's bare Grade x candidates have achieved in this year's examination?*

The task of the awarders is to answer questions of the above type and so arrive at judgements of the minimum marks required for each grade (called the grade boundaries) through the study of scripts and/or other work submitted for candidates alongside statistical and other information which is also available. Thus, the emphasis in the Board's awarding procedures is on the maintenance of standards from year to year. To this end, a reference year (usually the previous one) is established for each awarding meeting. This reference year provides the archive scripts and data for statistical comparisons which are used as the basis for the current year's awarding decisions.

The work of the awarding meeting assumes that the question papers present a fair test of the syllabus and that the marking has been properly conducted so that each candidate's mark in each component is the best available measure of his or her performance in that component and thus that each candidate's total mark (after any appropriate scaling has been completed) is the best available measure of his or her performance in the examination as a whole. The awarders recommend the minimum mark required for each of three Key grades: A, B and E, for each component of the examination separately. These recommendations must be based upon two equally important forms of evidence:

a. their professional judgements of how the quality of the work found in scripts or other work submitted by the current candidates compares with the quality of the work of previous candidates;

b. the statistical data which are available showing how the marks awarded in the current examination compare with those awarded in previous years.

It is not realistic to expect awarding meetings reliably to identify and accurately to correct major differences of awarding standard between different A Level examinations. Accordingly, if such major differences are suspected, the Board's procedures for dealing with them involve a more thorough investigation of all the relevant information than is possible within the constraints of an awarding meeting. Such investigations are carried out in response to SEAC scrutiny reports or Inter-Board research studies, at the request of the relevant *Standing Advisory Committee* or in response to any other information which raises serious questions about the standards of a particular examination. Their results are used as the basis of any necessary adjustment of standards at the following awarding meeting.

3. A-LEVEL AND AS AWARDING IN 1993

This section describes the roles of the various personnel involved in grade awarding meetings, the outcomes they are required to produce and the procedures to be used in 1993.

3.1 Personnel and Responsibilities

The Awarding Meeting will be attended by:

a. the Chairman of the Standing Advisory Committee (or his/her nominated alternate) together with at least two other members of the Standing Advisory Committee (i.e. a minimum of three SAC members including the Chairman of the awarding meeting are involved);

b. the Chief Examiner(s) for the subject (including, as appropriate, Senior Examiners who have primary responsibility for any component, Senior Coursework Moderators and so on.);

c. the Subject Officer.

A designated support officer and other members of the Senior Staff of Education Division will also attend the meeting to give any help or guidance, which they judge necessary. Such staff may also be called into the meeting at any time if a problem arises.

The three different sets of people required to attend awarding meetings - SAC members, Senior Examiners and the Subject Officer - each have a particular kind of expertise and information to contribute. When each has had an opportunity to do so a discussion of the evidence should

lead to recommendations about the placement of the key grade boundaries which represent the consensus of the meeting. In the event of minor disagreement when no evidence to resolve it can be found, a majority recommendation is acceptable. In the event of major dispute, when all the resources of the Grade Awarding Meeting have been exhausted and all the help and advice which can be gathered from Senior Staff and Research Officers has failed to resolve the difficulty, the responsibility for arriving at final recommendations rests with the SAC representatives.

The awarding meeting makes its recommendations to the Board. Provided that these recommendations are consistent with the Board's established expectations (see Section 3.6 Paragraph e., below) they will normally be accepted unchanged. On occasions when the recommended grade boundaries are not consistent with established expectations, the Board, which is represented for this purpose by the Secretary General, will consider the recommendations, the reasons for them and any other relevant matters before deciding upon the boundaries which should be used to award candidates' grades in the examination in question. However, grade boundaries different from those recommended by the Grade Awarding Panel will be adopted only with the express agreement of the Chairman of the Board's Education Committee.

3.2 Before the meeting

Before the Grade Awarding Meeting, each member of the Grade Awarding Panel will receive from the Subject Officer a set of materials to enable them to prepare for it. This will include the paper(s) and marking scheme(s), mark allocations and scaling factors for the current examination, together with the same, plus archive scripts, from the previous year's examination. The members of the Panel should study these in order both to remind themselves of the standards applied in the past and to make initial judgements about the relative difficulty of gaining marks in the two examinations.

The Chief Examiner will have initial ideas about possible grade boundaries, based upon his/her experience of the examination and its marking, together with reports from Assistant Examiners and discussions with Team Leaders, aided by a specially conducted scrutiny of scripts. These initial ideas, together with statistical information relating to the marks awarded, will be discussed with the Subject Officer well before the Grade Awarding Meeting. On the basis of this discussion, the Subject Officer will select scripts for use at the meeting.

3.3 The Outcomes of the Meeting

The main outcomes of the grade awarding meeting are the recommendations for the key grade boundaries (A/B, B/C and E/N) for each component. Once approved, these are combined,

by computer, in a prescribed way (described in Section 3.6) to give overall Subject boundaries for the examination as a whole. Should the recommended component boundaries imply changes in the cumulative percentages of candidates at the key Subject grade boundaries which are greater than the expected maxima established by the Board, then they must be accompanied by a written rationale prepared before the awarding meeting disperses. Further details concerning the written rationale are given in Section 3.7.

3.4 Initial Procedures

a. Chairman's Introduction

The Chairman should open the meeting by reminding the members of the task which has to be undertaken and outlining the procedures to be followed. The Chairman may also wish to draw attention to any ancillary materials which the Subject Officer may have provided such as any comparability or scrutiny reports.

b. Subject Officer's Introduction

The Subject Officer will then deal with domestic matters and remind the members about the details of the procedures and of the examination itself. These will include the mark allocations for each component and any scaling factors which have been applied in order to achieve the weightings set out in the syllabus. He or she will provide the members with any other information about the examination as a whole which it is appropriate to consider before the detailed work of the day begins. The Subject Officer will also give the meeting details about the characteristics of the entry for the current year and the previous reference year.

3.5 Establishing provisional key boundaries for the components

a. Order of components

The order in which the components are tackled is not prescribed except that externally marked written papers, where they exist, must be considered before all other types of component.

b. Senior Examiner's report

The Senior Examiner responsible for the first component begins its consideration by reporting on the way in which it has functioned. In particular, he or she should refer to any questions which were easier or harder for candidates to respond to than expected. If there are any questions in the current examination for which there were closely related counterparts in the reference year's examination, then the Senior Examiner should point out any perceived differences in the questions themselves which would have affected the ease with which candidates could score marks on such questions

compared with their previous counterparts. If the marking scheme used in the current year differs from the reference year in the proportion of marks awarded to a particular skill, set of skills or class of knowledge, then the probable effects of this upon the marks awarded should be explained. The essence of the Senior Examiner's report should not be his or her judgement of how the candidates have performed but, rather, the effect of specific features of particular questions (which either presented particular difficulties to the candidates or which encouraged particularly good responses), together with the current marking scheme, upon the ease with which candidates gained marks in the current examination.

c. Subject Officer's report

The Subject Officer then outlines the statistical evidence concerning the difficulty of obtaining marks in the first component in this year's examination compared with the same component in the previous year's examination (or the reference examination being used if this is not the immediately preceding one). The Subject Officer will compare the means and standard deviations of the marks for the component in the two years and draw out the implications for the difficulty of the component in the current year's examination.

d. The first boundary: E/N

The meeting then turns to the determination of the first component grade boundary: E/N. The Subject Officer informs the meeting of the mark which is statistically equivalent to the lowest mark in the previous year's Grade E for this component and of the starting range of marks which has therefore been chosen as likely to contain the E/N boundary on this component.

e. Scrutiny of Scripts

The members of the meeting now begin scrutinising scripts within the starting range with the purpose of identifying, in their professional judgement, which mark is attached to scripts which represent the same standard as those with the lowest component mark in Grade E in the previous year. Where sufficient scripts are available, it is helpful if each script is scrutinized by no more than two of the awarders. If this can be arranged, it ensures that a reasonable number of scripts are considered by the meeting as a whole and inhibits excessive discussion of individual, possibly unrepresentative, examples of candidates' work. For each script which they scrutinise, the awarders record the grade which they believe it to be worth on the pro-forma provided for this purpose. The essence of the process is comparative; the task is to judge this year s scripts, as responses to this year's question paper marked according to this year's marking scheme, using standards inferred from last year's

archive scripts, question paper and marking schemes for the same component. This is the case even for a new or revised syllabus although the comparison is then more difficult to make, being with a different syllabus in the same subject area rather than with the same component on a different occasion.

f. The initial provisional recommendation

Once each member is satisfied that he or she has reached a tentative conclusion about the location of the grade boundary, the Chairman calls the meeting to order. There then follows a discussion of the results of the script scrutiny and of the statistical evidence presented earlier by the Subject Officer. This discussion may lead to the adoption of a provisional recommendation for the location of the E/N boundary for the component. It is likely, however, that awarders will not all be able to agree on a single mark, but that they will be able to agree that all scripts above a certain mark are worthy of Grade E and that all scripts below some other mark are not. The establishment of such a borderline mark-range is all that is required at this stage in the procedure.

g. The other key grade boundaries: B/C and A/B

The procedures specified in d. to f. above are now repeated for the other two key boundaries B/C and A/B, in that order.

h. Review of provisional recommendations for the component

Once provisional recommendations for all three key boundaries (or provisional borderline mark-ranges) have been agreed, the results for the component as a whole should be reviewed. At this stage, checks upon the consistency of the qualitative judgements at the three grade boundaries will be made by the Subject Officer and discussed in the meeting. For example, comparing the details of the component grade boundaries with the previous year may show up anomalies between this year's provisional grade boundaries and the confirmed grade boundaries of last year; it would be difficult to justify placing the A/B grade boundary for a component at a higher mark than was chosen last year whilst, at the same time, placing the E/N grade boundary at a lower mark than last year unless the standard deviation of the marks had increased. Furthermore, as with subject boundaries (see Section 3.6, Paragraph e.) in examinations with entries of over 500 candidates, the cumulative percentages of candidates at the component grade boundaries for the current examination should not differ by more than a few percentage points from those in the previous year unless there is some independent reason to believe that the candidates, as a group, are more or less competent than last year's. Certainly, a large change in the cumulative percentage of candidates above, say, the B/C boundary which was unaccompanied by a corresponding change in the percentage above the A/B boundary would be

suspect. Considerations of this kind may lead to further modification and refinement of the provisionally agreed component grade boundaries. They also help an appropriate value to be chosen in those cases where a borderline mark-range, rather than a boundary mark, was agreed after the scrutiny of scripts.

l. The other components

The procedures set out in Paragraphs b. to h. above are then repeated for all other components for which scripts or other material evidence is available. The Subject Officer will provide detailed guidance on the interpretation of work produced in response to components other than externally marked written papers (for example, coursework, projects or orals).

(i) Coursework

For coursework components, where the work of each candidate is characteristically fairly extensive, it is not normally possible for the awarders to scrutinise several examples of work on each mark point and the conclusions which they reach from the scrutiny of candidates' work are therefore less reliable than is the case with externally marked written components. For coursework components, therefore, greater emphasis should be placed upon the statistical evidence than in the case of externally marked written papers. Where the requirements for a coursework component have been specified and the marks ascribed on the same basis as hitherto, there is no reason why the grade boundaries should alter much from year to year. Where requirements or marking criteria have changed, however, the statistically equivalent boundaries should normally be used.

(ii) Components where the work is ephemeral

For components where no candidates' work is available for scrutiny at all, it is necessary to receive a report from the Senior Examiner responsible for the particular component describing examples of the standard of work achieved by candidates. In such cases, the principal evidence which should be used to fix the grade boundaries is the statistical evidence. As with coursework, where the requirements for such a component have been specified and the marks ascribed on the same basis as hitherto, there is no reason why the grade boundaries should alter much from the previous year. Where requirements or marking criteria have changed, however, the statistically equivalent boundaries should normally be used

(iii) Objective Tests

The case of Objective Test components is a special one. Setting comparable grade boundaries on such tests by other than purely statistical means is generally accepted to be extremely difficult. It is not, therefore, part of the Board's procedures for the awarding meeting to set boundaries for Objective Test components. In examinations which use them, the process of determining subject boundaries is modified in such a way as to render judgements about Objective Test component boundaries unnecessary. Nevertheless, since the candidates' individual grades for the subject as a whole are arrived at by aggregating the marks for all components (including any Objective Test), performance in the Objective Test contributes in the normal way to each candidate's subject total and hence to the grade which they achieve.

(iv) Special overseas papers

The grading of these papers is done by statistically adjusting the grade boundaries set on the corresponding papers for home candidates. The awarding meeting is not concerned with this process.

3.6 Checking the Subject Boundaries

Once the three key boundaries have been provisionally determined for each component, it is necessary to check upon their effect for the subject as a whole. In particular, a check must be made that any changes from the previous year in the cumulative percentages of candidates at each of the key Subject grade boundaries do not exceed the Board's maximum expectations. To make this check, the provisional component boundaries must be combined to produce boundaries for the subject as a whole. There are two methods for doing this; both are used at each boundary and the lower one (which is exceeded by the marks of the greater number of candidates) is chosen.

a. <u>The percentile method</u>

For the E/N boundary, the Subject Officer first determines, from each unit distribution, the cumulative percentage of candidates whose marks exceed the provisional boundary mark on each component of the examination. The average (mean) of these percentages is then calculated (taking appropriate account of the weight given to each component in the syllabus) and the total mark for the subject as a whole which, as nearly as possible, is exceeded by this mean percentage of candidates is found. This total mark is one possible E/N boundary mark for the subject as a whole.

b. The addition method

In this method, the Subject Officer simply adds together the provisional component boundaries using the normal scaling factors for the components of the examination. The result is a second possible grade boundary for the subject as a whole.

c. Comparison

The Subject Officer then determines which of the two possible Subject E/N boundaries is the lower and this becomes the putative Subject E/N boundary.

d. The other key boundaries: B/C and A/B

The Subject Officer then repeats the procedures in Paragraphs a. to c. for the other two key boundaries B/C and A/B.

e. Comparison of statistical outcomes with the Board's expectations

For examinations attracting more than 500 candidates, the Board has established the maximum change in the cumulative percentage of candidates at each key Subject boundary which is to be expected in any one year. These are as shown below:

Grade	Expected Maximum annual change (cumulative percentage points)
A	1%
B	2%
E	3%

The Subject Officer now checks the cumulative percentage changes from the previous year's results which follow from the putative Subject boundaries determined in Paragraph c. above. (If the reference examination is not from the immediately preceding year, the comparison is with the year in which it was taken.) The purpose of this check is to see if the changes are within the expected maxima set out above. The action to be taken depending upon the result is prescribed in Paragraphs g. to i. below.

In 1993, the comparison of statistical outcomes is made more complex by the special phasing-in procedure used in 1992 for boundaries derived by the percentile method. For such boundaries (normally A/B and B/C), the cumulative percentage of candidates awarded the grade by the putative boundary should not be compared with the cumulative percentage actually awarded the grade in 1992. Rather, it must be compared with the cumulative percentage **which would have been awarded the grade if the phasing-in procedure had not been used for the boundary in question in 1992.** (These cumulative percentages were computed as part of the 1992 awarding procedure and will, therefore, be available from the records of the 1992 awarding

meetings.)

f. Small entry subjects

For examinations with fewer than 500 candidates, greater changes than those shown in Paragraph e. are likely to occur from year to year with greater frequency. As a result no limits are prescribed. However, the expected changes for examinations in this category are set out in Annex A to this paper. Awarding committees should bear these in mind when reviewing the statistical consequences of their recommendations; further action is not prescribed for small entry subjects but the general approach specified in Paragraphs g. to i. below should be followed as appropriate.

For small entry subjects covered by this section, the awarders' recommendations will normally be accepted on behalf of the Board by the Subject Officer, after consultation with the designated support officer for the meeting in question. However, either of these officers may refer the recommendations to the Secretary General (or his nominated alternate) if they judge it necessary.

g. If the cumulative percentage change is within the expected annual maximum at any grade

At any grade for which the cumulative percentage change for the Subject as a whole, based on the putative Subject grade boundary, is within the expected annual maximum, the provisional component boundaries can be adopted as final recommendations without further consideration and can be accepted on behalf of the Board by the Subject Officer.

h. If the cumulative percentage of candidates at a grade changes by more than the expected annual change

It is occasionally the case that a change in cumulative percentage exceeds the Board's established expectations for a statistical or technical reason. For example, in certain circumstances, the change in cumulative percentage for the subject as a whole can just exceed the established expectation even though the awarders recommend the statistically equivalent boundary for each component. Another example occurs in the case of subjects with common components where boundaries which meet the expectations simultaneously for each such subject cannot be found. In these circumstances, advice must be sought from the designated support officer for the meeting who, if the explanation is technical, will assist the Subject Officer in drawing up a written statement explaining the reasons for exceeding the Board's established expectations. Once this has been done, the provisional component boundaries can be adopted as firm recommendations and accepted on behalf of the Board by the Subject Officer. However, if they judge it necessary, either the Subject Officer or the

support officer may refer the recommendations and technical explanation to the Secretary General (or his nominated alternate) for approval.

In the majority of cases, where there is no technical explanation for a cumulative percentage change greater than the expected annual maximum, the provisional component boundaries must be reconsidered by repeating Procedures e., f. and h. from Section 3.5. It may be the case that the provisional boundary at the grade in question for one of the components, in particular, is doubtful because there was considerable divergence between the awarders' professional judgements of the sample scripts or because the provisional boundary was fixed an unusually large number of marks away from the statistically equivalent boundary established by the Subject Officer. In cases like this, it makes sense to begin the reconsideration with the particularly doubtful component. Otherwise, each component must be reconsidered at the boundary in question. If the reconsidered component boundaries lead to a putative Subject boundary which gives a change in cumulative percentage within the expected maximum, then they can be adopted as firm recommendations and accepted on behalf of the Board by the Subject Officer.

i. If the cumulative percentage change is greater than the expected annual change after the provisional component boundaries have been reconsidered

If, after reconsideration of the provisional component boundaries, the cumulative percentage change is still greater than the expected annual change, then the awarders must discuss the situation with the designated support officer for the meeting and/or with a member of the Senior Staff of Education Division. If, after that discussion, the awarders still wish to maintain their decisions, the reconsidered component boundaries go forward from the meeting as the recommended boundaries. In this case, a written rationale for the recommended component boundaries must be prepared **before the meeting disperses.**

3.7 The Written Rationale

If an awarding panel carries out a reconsideration of their provisional component boundaries but still wishes to make recommendations which produce greater than expected changes in the cumulative percentage of candidates at one or more grades, then they may do so. However, they are required, in these circumstances, to accompany those recommendations with a written rationale which is **prepared before the meeting disperses.** The written rationale will normally be prepared by the Subject Officer, acting as Secretary to the awarding meeting. The rationale should be signed by the Chairman to indicate that it represents the consensus view of the meeting. It is necessary for the written rationale to be detailed and specific. Bland *post hoc* justifications such as "the quality of the candidates' work is higher/lower this year" are not sufficient.

The written rationale must first address the issue of the relative difficulty of scoring marks in the current and previous year's examinations, supporting the view taken on this matter by explicit reference to the questions and marking schemes from the two examinations. It must also discuss the implications of the mark statistics for the relative difficulty of the current examination. The rationale should then turn to the performance of candidates at the recommended component boundaries and discuss in detail how this compares between the two years in terms of the features which are relevant to judgements of the quality of the work. From the scrutiny of candidates' scripts, there are likely to be insights about the nature of the work of the current year's candidates to which reference should be made. For example, in a physical science subject a deterioration in the capacity of candidates to understand and use algebraic expressions might be observed. Provided that the algebraic demands of the current papers were judged no greater than those of last year, then an effect of this kind could be quoted to support the view that an accompanying drop in the mean mark was the result of reduced performance by the candidates rather than a harder examination. As a result, lower percentages of candidates awarded the top grades could be expected. Other comments may also be made; for example, that one year's candidates are better/worse prepared than the other's. However, any such assertion **must be supported by evidence** in the form of relevant specific examples or a description of the relevant features of candidates' work from which this inference is drawn.

Preparing the rationale is likely to be time consuming and difficult and to be done at the end of a long, and possibly argumentative, day. However, recommendations for boundaries which produce greater than expected changes in the cumulative percentage of candidates at one or more grades must be accompanied by an appropriately detailed written rationale if they are to have any chance of being approved by the Board.

3.8 Interpolating the other Subject boundaries

Once the key subject boundaries have been established, the remaining subject boundaries will be obtained by interpolation/extrapolation by the Subject Officer following national guidelines. It is not **essential** for these interpolations to be made at the meeting because they will eventually be made by computer. However, the awarders may ask the Subject Officer to do so in order to get as complete a picture of the results of their deliberations as possible.

3.9 Recording the recommendations and implementing the results of the Meeting

Finally, at the end of the meeting, the recommendations of the awarders are recorded clearly on the standard form and a check made with the members that they are all agreed that this

written record is correct. The Chairman and Subject Officer then both sign the form to indicate that this is so.

In cases where the Subject Officer has not been able to accept the meeting's recommendations on behalf of the Board (see Section 3.6 Paragraph l.) approval for them is then sought. Immediately after the meeting, the Subject Officer will therefore take the recommendations, together with the written rationale to the Secretary General for approval. (In the absence of the Secretary General, the Deputy Secretary General will be consulted and, in his absence, the Head of Education Division.) In such cases, the Secretary General (or his alternate) determines the grade boundaries which will be used to award grades to candidates. Each case is considered on its merits but, in the absence of a clear rationale, boundaries which give statistical outcomes as similar as possible to those of the reference year are most likely to be used.

The completion of the form recording the awarding meeting's recommendations marks the formal end of the awarding meeting. Awarders may, if they wish, remain to hear the Secretary General's decision but this decision is final. If the awarders are concerned about the grade boundaries finally adopted, these concerns should be raised when the responsible Standing Advisory Committee reviews the outcomes of the examination at its Autumn meeting. At that time, the Standing Advisory Committee can request an investigation of the standards of the awards to be carried out by the Research and Statistics Group. The results of any such investigation will be used to inform the following year's awarding meeting.

In cases where the Subject Officer has been able to accept the meeting's recommendations on behalf of the Board (see Section 3.6 Paragraphs f., g. and h.), he or she **immediately** informs the Secretary General, his deputy or the Head of Education Division, as appropriate, that this is the case.

When an approved set of grade boundaries is available, they will be entered into the Board's computer and the process which selects candidates whose work is to be borderline re-marked will be initiated.

4. CONCLUDING REMARKS

There are sound educational reasons why grade awarding should involve both qualitative judgements of candidates' scripts and a consideration of statistical information. To adopt a purely statistical approach and assign predetermined proportions of candidates to particular grades, year after year, cannot be justified. The overall attainment of the candidates entered for a given examination may vary slightly from year to year, perhaps as a result of different

teaching and learning conditions or a change in the types of centre from which the candidates are drawn. If grades are awarded to identical proportions of candidates every year, important long-term changes in levels of attainment might go undetected.

On the other hand, any such changes are likely to be small between two successive years so it is reasonable not to expect large changes in the proportions of candidates awarded each grade in any one year. Large changes in the examination statistics are most likely to arise in a single year because the examination under consideration is more or less demanding or has been marked more or less severely than hitherto. The evidence suggests that, in these circumstances, judgements based solely upon the perusal of candidates' scripts are unlikely to make sufficient allowance for the change in the examination and/or its marking. (Incidentally, it is changes of this sort which make it unjustifiable simply to set the grade boundaries at the same mark points every year.) By requiring Grade Awarding Panels to consider both their own professional judgements of the quality of candidates' work and the statistical evidence which is available, the Board believes that the levels of attainment required for the award of grades will be fair and consistent from year to year.

16th June 1993

NEWAPP2

ANNEX A

EXPECTED MAXIMUM ANNUAL CHANGES IN CUMULATIVE GRADE PERCENTAGES

FOR SMALL ENTRY SUBJECTS

300 - 500 candidates

Grade	Expected maximum annual change (cumulative percentage points)
A	2%
B	3%
E	4%

200 - 300 candidates

Grade	Expected maximum annual change (cumulative percentage points)
A	3%
B	4%
E	5%

100 - 200 candidates

Grade	Expected maximum annual change (cumulative percentage points)
A	4%
B	5%
E	7%

A LEVEL AND AS EXAMINATION AWARDING PRO-FORMA

SUBJECT .....

NAME OF AWARDER .....

YEAR .....

Component	Centre	Candidate	Mark	Grade	Comments

ARRIVING AT SUBJECT GRADE BOUNDARIES

For each key grade boundary, the lowest mark in grade is calculated by two different methods; then the method which favours the candidates, i.e. gives the lower mark, is used operationally. In Method 1, marks (scaled if necessary) are added together; Method 2 involves aggregating percentages.

Please note that the marks and percentages in the examples below are hypothetical. They should not be construed as the numbers to be expected in practice.

Method 1 - the addition method

This method is essentially the same as that used to obtain each candidate's mark for the subject. The one small difference concerns the treatment of fractions of a mark which can arise through scaling: for an individual candidate, a fraction is rounded up; in calculating a boundary, a fraction is rounded down. Both actions are designed so as not to disadvantage the candidates.

Let us consider a 3 component examination with the following characteristics:

	Paper 1	Paper 2	Paper 3
Maximum unscaled marks available	50	75	100
% of total scaled marks available	20%	30%	50%

To obtain the correct percentage of marks specified for each paper, the unscaled marks of Paper 3 have to be multiplied by a scaling factor of 1.25.

	Paper 1	Paper 2	Paper 3
Maximum unscaled marks available	50	75	100
x scaling factor	x1	x1	x1.25
Maximum scaled marks available	50	75	125
% of total scaled marks available	20%	30%	50%

The above example is set out on page 5. Let us assume that the Grade Awarding Meeting sets the A/B boundary – the minimum unscaled mark for Grade A – of Papers 1, 2 and 3 at 41, 62 and 82 marks respectively. To obtain the subject grade boundary by Method 1, we first scale the marks, add the scaled marks together, and finally round down any fraction:

$$(41 \times 1) + (62 \times 1) + (82 \times 1.25)$$

$$41 + 62 + 102.5 = 205.5$$

fraction rounded down: 205

So Method 1 gives 205 as the minimum aggregate scaled mark on the Subject Mark Distribution for Grade A.

Similar calculations by Method 1 are shown on page 6 for the other key boundaries, B/C and E/N.

UNIT MARK DISTRIBUTION

(Paper 1)

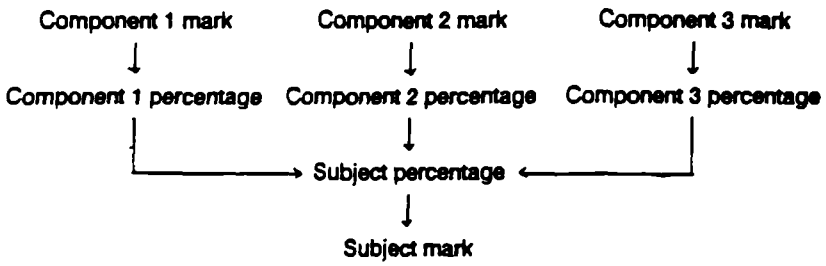
mark	number	cumulative number	cumulative percentage
50	2	(2)	0.42%
•	•	•	•
•	•	•	•
•	•	•	•
42	7	(67)	13.96%
41	5	(72)	15.00%
40	12	(84)	17.50%
39	17	(101)	21.04%
•	•	•	•
•	•	•	•
•	•	•	•
•	•	•	•
•	•	•	•
•	•	•	•
0	•	(480)	100.00%

SUBJECT MARK DISTRIBUTION

mark	number	cumulative number	cumulative percentage
250	0	(0)	0.00%
•	•	•	•
•	•	•	•
•	•	•	•
198	3	(52)	10.83%
197	5	(57)	11.87%
196	0	(57)	11.87%
195	8	(65)	13.54%
194	12	(77)	16.04%
•	•	•	•
•	•	•	•
•	•	•	•
•	•	•	•
•	•	•	•
•	•	•	•
0	•	(480)	100.00%

Method 2 - the percentile method

For a 3-component examination, the steps in Method 2 are these:



We describe below what is entailed in each arrow of the diagram above.

In our example, the lowest mark for Grade A on Paper 1 is 41. From the Unit Mark Distribution (page 2) we find that a mark of 41 puts 15% of the candidates in Grade A.

Corresponding marks for this boundary – and hence percentages of candidates – can be found for the other components. In the example on page 5 there are 13% Grade As for Paper 2 and 10% for Paper 3.

In calculating the overall percentage of candidates to be given Grade A, we take into account the percentage of candidates worthy of Grade A in each component and also the percentage of the total marks allocated to each component. The calculation is shown (page 5) which indicates that 11.90% of Grade A candidates overall is required. So we turn to the Subject Mark Distribution of our example (the relevant portion is shown on page 2) and in the column showing the cumulative percentages of candidates we look for the figure nearest to 11.90%, which in this case is 11.87%. In fact, 11.87% happens to appear twice – against marks 197 and 196. This occurs because there are actually no candidates with an aggregate scaled mark of 196 – something that may happen in a subject with a small number of entries. The boundary is deemed to fall immediately below the **lowest** mark to which the cumulative percentage relates. So, in this example, Method 2 gives 196 as the minimum aggregated scaled mark for Grade A.

For the B/C boundary, let us suppose that the lowest marks for Grade B in Papers 1, 2 and 3 are 35, 49 and 63 respectively, and that these three marks give rise to 32%, 31% and 28% of candidates with Grade B and above in the three papers. By calculation, we find that 29.70% of candidates should get Grade B and above in the subject overall. The mark for which the cumulative percentage is nearest to 29.70% is found on the Subject Mark Distribution and might be, say, 156. So by Method 2 the B/C boundary is located immediately below 156.

A similar procedure for the E/N boundary might give 90.70% of candidates overall with Grade E and above; and from the Subject Mark Distribution this might lead to the E/N boundary coming immediately below a mark of 64 by Method 2 (see page 6).

The Operational Boundaries

Assuming that the recommendations of the Grade Awarding Meeting are accepted, then the operational grade boundaries from our example are as follows:

At the A/B boundary, Method 1 gives a boundary mark of 205, and Method 2 gives 196. Method 2 (the percentile method) provides the lower mark. So the A/B boundary is drawn immediately below 196 on the Subject Mark Distribution.

Method 2 also gives the lower mark at the B/C boundary: 156, compared with 162. The B/C boundary thus comes immediately below 156.

Turning to the E/N boundary: Method 1 yields a mark of 60; Method 2 gives 64. So Method 1 (the addition method) prevails: the E/N boundary is set immediately below 60.

Other Assessment Patterns

The example given is for an Assessment Pattern of 3 papers; a similar series of calculations can be performed for examinations with 2 or 4 papers.

Syllabuses with an Objective Test

The Grade Awarding Meeting is not called upon to make judgements about grade boundaries on Objective Tests. However, it is nonetheless necessary for a mark to be identified for each key grade boundary on Objective Tests. How this is done is shown on pages 7 and 8. The example given is for a three component examination, Paper 1 of which is an Objective Test. Let us suppose the Meeting fixes the minimum mark acceptable for Grade A on Papers 2 and 3 at 62 and 82 marks respectively.

It is easier to explain the procedure with an Objective Test if we look first at Method 2: the percentile method. So that is where we shall begin. Reading off the Unit Mark Distributions might show us that 13% and 10% of candidates would obtain Grade A in Papers 2 and 3 respectively. The calculation (page 7) of the percentage of candidates who achieve Grade A in the overall subject reflects the 30% and 50% of the maximum subject mark allocated to Papers 2 and 3, and works out at 11.10%. So the recommended Grade A boundary by Method 2 falls immediately below the aggregate scaled mark which gives a percentage of candidates nearest to 11.10%: 198 in the example.

Method 1 requires a mark from Paper 1 to go in the calculation. In the course of working out Method 2, we calculated the overall percentage of Grade A candidates from the results of Papers 2 and 3: in this case, 11.10%. From the Unit Mark Distribution for Paper 1, the Objective Test, we find the unscaled mark which is nearest to that percentage. Suppose this mark is 43. We then use this in the calculation of the subject grade boundary, as described earlier.

And finally, we see which of the two methods yields the lower mark – Method 2 in this instance – and this is the recommended boundary.

This arrangement does not mean that a candidate's performance in the Objective Test has no effect on his or her grade; it simply means that the percentage of candidates achieving Grade A in Paper 1 is taken to be the same as that which does so in the subject as a whole when determined by the boundary decisions taken about Papers 2 and 3.

Similar decisions and subsequent calculations (which are shown on page 8) lead to the subject grade boundaries for B/C and E/N.

Objective Tests shared by more than one subject

When an objective test component is a unit in more than one subject, the process of finding a mark for use in the addition method (Method 1) is slightly more complex. This is because the different subjects concerned may produce marginally different estimates of the objective test boundary mark. Where this occurs, the mark used as the objective test boundary for the grade in question should be the mean of the different estimates, weighted to reflect the number of candidates entered for each subject. The support officer will compute this boundary mark, if required.

Obtaining Subject Grade Boundaries: an example

	Paper 1 Externally-assessed Written	Paper 2 Externally-assessed Written	Paper 3 Centre-assessed Coursework	Overall Subject
Component boundaries to be fixed judgementally by Grade Awarding Meeting	A/B B/C E/N	A/B B/C E/N	A/B B/C E/N	
Maximum unscaled marks available	*50 marks*	*75 marks*	*100 marks*	
Scaling factor	*×1*	*×1*	*×1.25*	
Maximum scaled marks available	*50*	*75*	*125*	*250 marks*
% of total scaled marks available	20%	30%	50%	100%
A/B Boundary				
Minimum unscaled mark for Grade A on Papers 1, 2 and 3 fixed judgementally *for example*	*41 marks*	*62 marks*	*82 marks*	
Method 1 adding marks				
Calculation of overall subject mark (fractions rounded down)	*(41 × 1)* +	*(62 × 1)* +	*(82 × 1.25)* =	*205 marks*
Method 2 aggregating percentages				
On each paper: minimum unscaled mark for Grade A, fixed judgementally, gives rise (on Unit Mark Distribution) to % of candidates with Grade A	15%	13%	10%	
Calculation of overall % of Grade A candidates	$\frac{(15 \times 20)}{20 + 30 + 50}$ +	$\frac{(13 \times 30)}{20 + 30 + 50}$ +	$\frac{(10 \times 50)}{20 + 30 + 50}$ % =	11.90%
On Subject Mark Distribution: scaled mark whose cumulative % is nearest to calculated value				*196 marks*

Method 2 (196 marks) gives lower boundary than Method 1, (205 marks), so operational boundary is by Method 2

APPENDIX 5.2

B/C Boundary

Minimum unscaled mark for Grade B on Papers 1, 2 and 3 fixed judgementally *for example* — *35 marks* — *49 marks* — *63 marks*

Method 1 adding marks

Calculation of overall subject mark (fractions rounded down): *(35 × 1)* + *(49 × 1)* + *(63 × 1.25)* = *162 marks*

Method 2 aggregating percentages

On each paper: minimum unscaled mark for Grade B, fixed judgementally, gives rise (on Unit Mark Distribution) to % of candidates with Grade B and above: 32% | 31% | 28%

Calculation of overall % of candidates with Grade B and above: $\frac{(32 \times 20) + (31 \times 30) + (28 \times 50)}{20 + 30 + 50}$ % = 29 70%

On Subject Mark Distribution: scaled mark whose cumulative % is nearest to calculated value: *156 marks*

Operational boundary is by Method 2

E/N Boundary

Minimum unscaled mark for Grade E on Papers 1, 2 and 3 fixed judgementally *for example* — *12 marks* — *18 marks* — *24 marks*

Method 1 adding marks

Calculation of overall subject mark (fractions rounded down): *(12 × 1)* + *(18 × 1)* + *(24 × 1.25)* = *60 marks*

Method 2 aggregating percentages

On each paper: minimum unscaled mark for Grade E, fixed judgementally, gives rise (on Unit Mark Distribution) to % of candidates with Grade E and above: 93% | 92% | 89%

Calculation of overall % of candidates with Grade E and above: $\frac{(93 \times 20) + (92 \times 30) + (89 \times 50)}{20 + 30 + 50}$ % = 90.70%

On Subject Mark Distribution: scaled mark whose cumulative % is nearest to calculated value: *64 marks*

Operational boundary is by Method 1

Obtaining Subject Grade Boundaries: an example containing an Objective Test

	Paper 1 Externally-assessed Objective test	Paper 2 Externally-assessed Written	Paper 3 Centre-assessed Coursework	Overall Subject
Component boundaries to be fixed judgementally by Grade Awarding Meeting	none	A/B B/C E/N	A/B B/C E/N	
Maximum unscaled marks available	*50 marks*	*75 marks*	*100*	
Scaling factor	*×1*	*×1*	*×1.25*	
Maximum scaled marks available	*50*	*75*	*125*	*250 marks*
% of total scaled marks available	20%	30%	50%	100%
A/B Boundary				
Minimum unscaled mark for Grade A on Papers 2 and 3 fixed judgementally *for example*		*62 marks*	*82 marks*	
Method 1 adding marks				
Mark on Objective Test nearest to % of candidates with Grade A as calculated from Papers 2 and 3	*43 marks*			
Calculation of overall subject mark (fractions rounded down)	*(43 × 1)* +	*(62 × 1)* +	*(82 × 1.25)* =	*207 marks*
Method 2 aggregating percentages				
On each paper: minimum unscaled mark for Grade A, fixed judgementally on Papers 2 and 3, gives rise (on Unit Mark Distribution) to % of candidates with Grade A		13%	10%	
Calculation of overall % of Grade A candidates		$\frac{(13 \times 30) + (10 \times 50)}{30 + 50}$ % =		11.10%
On Subject Mark Distribution: scaled mark (which includes Paper 1) whose cumulative % is nearest to calculated value				*198 marks*

operational boundary is by Method 2

B/C Boundary

Minimum unscaled mark for Grade B on Papers 2 and 3 fixed judgementally *for example* — *49 marks* — *63 marks*

Method 1 adding marks

Mark on Objective Test nearest to % of candidates with Grade B and above as calculated from Papers 2 and 3: *36 marks*

Calculation of overall subject mark (fractions rounded down): *(36 × 1)* + *(49 × 1)* + *(63 × 1.25)* = *163 marks*

Method 2 aggregating percentages

On each paper: minimum unscaled mark for Grade B, fixed judgementally on Papers 2 and 3, gives rise (on Unit Mark Distribution) to % of candidates with Grade B and above: 31% — 28%

Calculation of overall % of candidates with Grade B and above: $\frac{(31 \times 30) + (28 \times 50)}{30 + 50}$ % = 29.10%

On Subject Mark Distribution: scaled mark whose cumulative % is nearest to calculated value: *158 marks*

operational boundary is by Method 2

E/N Boundary

Minimum unscaled mark for Grade E on Papers 2 and 3 fixed judgementally *for example* — *20 marks* — *25 marks*

Method 1 adding marks

Mark on Objective Test nearest to % of candidates with Grade E and above as calculated from Papers 2 and 3: *11 marks*

Calculation of overall subject mark (fractions rounded down): *(11 × 1)* + *(20 × 1)* + *(25 × 1.25)* = *62 marks*

Method 2 aggregating percentages

On each paper: minimum unscaled mark for Grade E, fixed judgementally on Papers 2 and 3, gives rise (on Unit Mark Distribution) to % of candidates with Grade E and above: 92% — 89%

Calculation of overall % of candidates with Grade E and above: $\frac{(92 \times 30) + (89 \times 50)}{30 + 50}$ % = 90.10%

On Subject Mark Distribution whose: scaled mark cumulative % is nearest to calculated value: *64 marks*

operational boundary is by Method 1

APPENDIX 8.2

APPENDIX 5

EXPECTED MAXIMUM ANNUAL CHANGES IN CUMULATIVE GRADE PERCENTAGES FOR SMALL ENTRY SUBJECTS

300 - 500 candidates

Grade	Expected maximum annual change (cumulative percentage points)
A	2%
B	3%
E	4%

200 - 300 candidates

Grade	Expected maximum annual change (cumulative percentage points)
A	3%
B	4%
E	5%

100 - 200 candidates

Grade	Expected maximum annual change (cumulative percentage points)
A	4%
B	5%
E	7%

APPENDIX 6

NATIONAL GUIDELINES ON THE JUDGEMENTAL AND ARITHMETICAL FIXING OF GRADE BOUNDARIES AT A-LEVEL

It is agreed nationally that grading standards in 'A' Level examinations will be established in the following manner:

(i) The A/B grade borderline should be established on the basis of Boards' existing practice, including examiners' judgement of quality.

(ii) The B/C and E/N grade borderlines should be established by reference to examiners' judgements of quality and using methods to ensure fairness and reliability at these crucial points.

(iii) The mark range between the B/C and E/N grade borderlines should be divided into three equal intervals and these assigned respectively to grades C, D and E, with the same interval also assigned to grade N.

How this is done in the AEB is described below.

(a) The A/B, B/C and E/N grade boundaries will be fixed on the basis of grade awarders' judgements.

(b) The C/D and D/E grade boundaries will be fixed by equipartitioning the mark range between the lowest mark for grade B and lowest mark for grade E. (If the number of marks is not exactly divisible by three, the following will apply:

- where there is a remainder of one, after the B/C to E/N mark range is divided by three, then the one extra mark is added to the grade C range;
- where there is a remainder of two, one extra mark is added to both the grade C and grade D ranges. Examples of these cases are given below.)

(c) The N/U grade boundary will be fixed by extrapolating from the E/N grade boundary so that the mark range for grade N is the same as that for grade E.

Example 1

Consider an examination with a maximum mark allocation of 100 marks. The grade boundaries would be arrived at as described below.

1. To maintain the standard set in previous years, the awarders have decided that 70 should be the lowest mark for grade A.

2. To maintain the standard set in previous years, the awarders have decided that 60 should be the lowest mark for grade B.

3. To maintain the standard set in previous years, the awarders have decided that 41 should be the lowest mark for grade E.

4. The mark range between the B/C and E/N grade boundaries is 60 - 41 = 19 marks.

5. Since the mark ranges for grades C, D and E have to be equal, the C/D and D/E grade boundaries are arrived at in this case as follows.

   (i) The mark range for each of these three grades is $\frac{19}{3}$ = 6, remainder 1.

   (ii) Now, to give the candidates the maximum possible benefit, the extra mark is added to the grade C range, giving mark ranges of 7, 6 and 6 for grades C, D and E respectively.

   (iii) The lowest mark for grade C is 60 - 7 = 53 and for grade D is 53 - 6 = 47.

6. Since the mark range for grade N has to be the same as that for grade E, the lowest mark for grade N is obtained by subtracting 6 (the mark range for grade E) from 41, the lowest mark for grade E, giving 35.

The results are summarised in tabular form below.

Grade	Lowest Mark	Mark Range	Method
A	70	100 - 70	Inspection
B	60	69 - 60	Inspection
C	53	59 - 53	Interpolation
D	47	52 - 47	Interpolation
E	41	46 - 41	Inspection
N	35	40 - 35	Extrapolation
U	0	34 - 0	-

Example 2

Consider an examination with a maximum mark allocation of 100 marks. The grade boundaries would be arrived at as described below.

1. To maintain the standard set in previous years, the awarders have decided that 69 should be the lowest mark for grade A.

2. To maintain the standard set in previous years, the awarders have decided that 61 should be the lowest mark for grade B.

3. To maintain the standard set in previous years, the awarders have decided that 41 should be the lowest mark for grade E.

4. The mark range between the B/C and E/N grade boundaries is 61 - 41 = 20 marks.

5. Since the mark ranges for grades C, D and E have to be equal, the C/D and D/E grade boundaries are arrived at in this case as follows.

   (i) The mark range for each of these three grades is $\frac{20}{3}$ = 6, remainder 2.

   (ii) Now, to give the candidates the maximum possible benefit, the extra marks are added to the grade C and D ranges, giving mark ranges of 7, 7 and 6 for grades C, D and E respectively.

   (iii) The lowest mark for grade C is 61 - 7 = 54 and for grade D is 54 - 7 = 47.

6. Since the mark range for grade N has to be the same as that for grade E, the lowest mark for grade N is obtained by subtracting 6 (the mark range for grade E) from 41, the lowest mark for grade E, giving 35.

The results are summarised in tabular form below.

Grade	Lowest Mark	Mark Range	Method
A	69	100 - 69	Inspection
B	61	68 - 61	Inspection
C	54	60 - 54	Interpolation
D	47	53 - 47	Interpolation
E	41	46 - 41	Inspection
N	35	40 - 35	Extrapolation
U	0	34 - 0	-

SUBJECTS (OPTIONS) SHARING COMPONENTS

Component Distributions

For components which are shared by two or more normal enterable A-level subjects, component mark distributions are required in addition to unit mark distributions. Subject Officers who are responsible in 1993 for the awarding of any of the components listed below must, therefore, request such distributions from Computing Services Division. These component distributions are needed to determine statistically equivalent boundaries for the components (and thus units) in question and are, therefore, required at the same time as the unit distributions - immediately before the awarding meeting.

In June 1993, the following components are shared by two or more normal enterable A-level subjects:

Component Code	Subjects
A/MATH/1	0632 Pure Mathematics 0636 Pure and Applied Mathematics 0646 Pure Mathematics and Statistics
A/MATH/3	0602 Applied Mathematics 0636 Pure and Applied Mathematics 0649 Applied Mathematics and Statistics
A/MATH/9	0641 Statistics 0646 Pure Mathematics and Statistics 0649 Applied Mathematics and Statistics
A/BIOL/1, A/BIOL/2	0607 Biology 0677 Biology
A/SOCLGY/1	0639 Sociology 0664 Sociology (with CWK)
A/HIST/01	0630AF/AG/AH/AK/AM History 0673A History (Alternative)
A/HIST/02	0630BF/BG/BH/BK/BM History 0673B History (Alternative)
A/HIST/03	0630CF/CG/CH/CK/CM History 0673C History (Alternative)

	0673D History (Alternative)
A/EC HIST/1	0673E History (Alternative)
	0620 Economic and Social History
0630/2/05	0630AF/BF/CF/DF History
0630/2/06	0630AG/BG/CG/DG History
0630/2/07	0630AH/BH/CH/DH History
0630/2/08	0630AK/BK/CK/DK History
0630/2/09	0630AM/BM/CM/DM History
0635/1, 0635/2	0635A Physics
	0635B Physics
0637/01	0637A/B/C/D Religious Studies
0637/02	0637A/E/F/G Religious Studies
0637/03	0637B/E/H/K Religious Studies
0637/04	0637C/F/H/M Religious Studies
0637/05	0637D/G/K/M Religious Studies
0654/1, 0654/2, 0654/3	0654A Chemistry
	0654B Chemistry
0673/2, 0673/3	0673A/B/C/D/E History (Alternative)

The component distributions should also be used **in place of the unit distributions** for the purposes of reviewing the awarders' provisional component boundary recommendations (see Section 3.4h of the main procedure file).

Checking Subject Boundaries

For a syllabus containing option subjects, the following procedure is used to calculate and check the subject boundaries.

For the option with the most candidates and for any other options with over 500 candidates, the option grade boundaries are calculated in the usual way using the addition method or percentile method, whichever is more favourable. In each case, the resulting cumulative percentage is compared with that for the grade in question in the appropriate *option* distribution for the reference year. The normal limits and procedures set out in Section 3.5 apply to each such option. Thus, if any option with more than 500 candidates produces results outside the normal statistical expectations, re-consideration of the component boundaries is indicated.

However, it is possible for the examination data to be such that no component boundaries can be found so that every option has results within the normal limits. In these circumstances, the Subject Officer and

designated support officer for the meeting will draw up a written technical explanation of the outcome in the discrepant option(s). The component boundaries recommended should produce statistical outcomes within the normal expectations for the option taken by the largest number of candidates. A written rationale is required from the awarders if they wish to recommend component boundaries which do not meet this latter condition.

For options with fewer than 500 candidates (except the largest option, if this has fewer than 500), the option boundaries are calculated from those for the largest option, taking into account any differences in the unit boundaries. This method is preferable to, and therefore replaces, the use of the usual percentile method for these small options. No subsequent statistical check of the option boundaries is needed for options with fewer than 500 candidates.

In order to take account of the differences between the unit boundaries of the small option and the unit boundaries of the largest one, the addition method boundaries for each option must be computed. The addition method boundary for the small option is then subtracted from that of the largest option to find the difference, d. This difference, d, is then subtracted from the final boundary (addition or percentile, as the case may be) for the largest option.

Example

Subject 0999 has a number of options. For Paper 1, candidates take 0999/1/01, 0999/1/02 or 0999/1/03, while for Paper 2 candidates take 0999/2/01 or 0999/2/02. There is no scaling on Paper 1 but Paper 2 has a scaling factor of x2. The largest option (0999A) consists of 0999/1/01 + 0999/2/01, while option 0999E, consisting of 0999/1/03 + 0999/2/02, is an option with less than 500 candidates.

The table shows the recommendations for the component boundaries for these two options, together with the final boundaries.

	Paper 1		Paper 2		Addition	Percentile	Final
	boundary	scaling	boundary	scaling	boundary	boundary	boundary
0999A Largest Option	80	x1	70	x2	220	210	210
0999E Small Option	77	x1	69	x2	215	not needed	205*

Difference between addition boundaries, d = 220 - 215 = 5

*Final boundary for 0999E = 210 - d = 210 - 5 = 205

APPENDIX 8

SUBJECTS WITH ALTERNATIVE OVERSEAS PAPERS

The table below shows those subjects which, for the first time in 1993, also exist in special overseas versions created by setting an alternative paper to replace one or more papers sat by the home candidates.

Component for which an alternative paper will be set	Subjects affected	New subject codes for overseas cands
0600/1 0600/2	0600 Accounting	0600A
0618/1	0618 Economics	0618A
	0618S Economics	0618AS
0625/1	0625 Law	0625A
0635/1	0635A Physics	0635CA
	0635AS Physics	0635DA
	0635B Physics	0635CB
	0635BS Physics	0635DB
0655/1	0655 Business Studies	0655A
A/BIOL/1	0677 Biology	0677A
	0677S Biology	0677AS
A/MATH/1 A/MATH/9	0632 Pure Maths	0632A
	0632S Pure Maths	0632AS
	0636 Pure and Applied Maths	0636A
	0636S Pure and Applied Maths	0636AS
	0641 Statistics	0641A
	0646 Pure Maths and Stats	0646A
	0649 Applied Maths and Stats	0649A
	0690 Alternative Award	0690A
	0691 Alternative Award	0691A

The grading of these special overseas versions is described in this appendix.

GRADING THE NORMAL HOME SUBJECT

The grading of the normal home candidate subjects in the above list follows normal procedures, except that it is necessary to use special statistical reference data in 1993 (the first year of operation of the special overseas versions). **The distributions from the 1992 awarding meetings must not be used for statistical reference purposes in 1993 in the subjects listed above.** Instead, special unit, subject (and, where necessary, component) 1992 mark distributions, excluding the relevant overseas candidates, have been provided to Subject Officers for these

purposes by the Research and Statistics Group. These special distributions should be used in place of the data from the 1992 awarding meeting to determine the statistically equivalent component boundaries, to monitor the awarders' provisional component grade boundaries and for checking that any changes in overall subject outcomes are within the Board's established expectations.

GRADING THE SPECIAL OVERSEAS VERSION OF THE SUBJECT

All the above subjects except Accounting (0600A)

In all the above subjects except Accounting, the awarding committee will not be asked to make grade boundary decisions on the alternative papers for overseas candidates. The awarding committee will set grade boundaries only for those subjects and components that exclude overseas candidates. The grade boundary decisions on the common papers will also apply to the overseas components (e.g. the grade boundary decisions for 0618/2 will also apply to 0618A/2). The grade boundaries on the alternative papers will be calculated using structural regression to modify the grade boundaries on the home papers by taking account of the performances of the two groups of candidates on the common paper(s). The technical details of the method are described below using 0618 Economics as an example.

Let m_i = grade boundary i on 0618/1 (set by awarders)
M = mean mark on 0618/1
s_m = standard deviation on 0618/1

x_i = grade boundary i on 0618A/1 (to be calculated)
X = mean mark on 0618A/1
s_x = standard deviation on 0618A/1

c_i = composite grade boundary i on common papers 0618/2 and 0618/3
C_m = composite mean mark on 0618/2, 0618/3
s_{cm} = composite standard deviation on 0618/2, 0618/3
C_x = composite mean mark on 0618A/2 and 0618A/3
s_{cx} = composite standard deviation on 0618A/2 0618A/3

The z-scores of 0618/1 and 0618A/1 are set equal to the z-scores of the composites of papers 2 and 3 thus:

$$\frac{c_i - C_m}{s_{cm}} = \frac{m_i - M}{s_m} \quad \text{and} \quad \frac{c_i - C_x}{s_{cx}} = \frac{x_i - X}{s_x}$$

Which equations, when combined and re-arranged to eliminate c_i give the following formula for x_i the grade i boundary on the alternative paper:

$$x_i = m_i \cdot \frac{s_{cm} \cdot s_x}{s_m \cdot s_{cx}} + \left[C_m - C_x - M \cdot \frac{s_{cm}}{s_m} \right] \cdot \frac{s_x}{s_{cx}} + X$$

This equation can be written in the form:

$$x_i = m_i \cdot K_1 + K_2$$

where K_1 and K_2 are constants for any particular pair of home and overseas examinations.

Grade boundaries for the alternative paper will be calculated, using this equation, by the support officer after the awarding meeting for the normal home subject is over. The values of K_1 and K_2 for each pair of examinations will be provided by the Research and Statistics Group. The results of this process will be checked by reference to the performance of the relevant overseas candidates in 1992. For this purpose, distributions of their marks in 1992 will be provided to the support officer by the Research and Statistics Group. If no written rationale was required (or a rationale was written but not accepted) for the normal home subject, then the alternative paper boundaries will be adjusted, if necessary, to bring any changes in the outcomes for the overseas version of the subject within the Board's normal expectations. If a rationale was approved for the normal home subject, then adjustments will be made, if necessary, to reflect the consequences of the home subject recommendations in the overseas version

Accounting (0600A)

For Accounting (0600A) there are no papers in common with the normal home subject, 0600. Structural Regression can not, therefore, be applied in this case and the awarding committee will need to make grade boundary decisions for the alternative papers. The awarding of this subject will, therefore, follow standard procedures with distributions of the relevant overseas candidates' marks in 1992 being used for all statistical reference purposes. These distributions will be provided to the Subject Officer by the Research and Statistics Group.

Associated Examining Board

General Certificate of Education

Component Grade Boundaries

June / November 19

Syllabus Code .....

Syllabus Name .....

Well before the meeting complete the component codes and the grades you know will be required (in most cases the grades will be A, B and E). If possible, check the component codes against the Listing f AEB Syllabuses, Subjects and Components.

Component Codes	Maximum possible raw mark	Grades and lowest raw mark					
		Grade	Mark	Grade	Mark	Grade	Mark
_____/ ___							
_____/ ___							
_____/ ___							
_____/ ___							
_____/ ___							
_____/ ___							
_____/ ___							

After full consideration, in accordance with the relevant procedures of the Associated Examining Board, the above grade boundaries are recommended.

Signed ..... (Chairman) Date .....

Signed ..... (Subject Officer) Date .....

Approved ..... (*If rationale required*) Date .....

APPENDIX 10

SUBJECT : _____ **YEAR :** ____

ENTRY DETAILS	Reference Year 199_	199_
Private and External		
F.E.		
Overseas		
Schools		
Male		
Female		
Total Entry		

STATISTICS

Paper 1	Reference Year	Statistical Equivalence	Recommended Boundaries
Mean/S.D.			
Raw Mark Max/ Scaling Factor			
% weighting			
A/B boundary / cum. %			
B/C boundary / cum. %			
E/N boundary / cum. %			

Paper 2

Mean/S.D.			
Raw Mark Max/ Scaling Factor			
% weighting			
A/B boundary / cum. %			
B/C boundary / cum. %			
E/N boundary / cum. %			

Paper 3

Mean/S.D.			
Raw Mark Max/ Scaling Factor			
% weighting			
A/B boundary / cum. %			
B/C boundary / cum. %			
E/N boundary / cum. %			

Paper 4

Mean/S.D.			
Raw Mark Max/ Scaling Factor			
% weighting			
A/B boundary / cum. %			
B/C boundary / cum. %			
E/N boundary / cum. %			

Paper 5

Mean/S.D.			
Raw Mark Max/ Scaling Factor			
% weighting			
A/B boundary / cum. %			
B/C boundary / cum. %			
E/N boundary / cum. %			

	Addition Method	Percentile Method
A/B boundary / cum. %		
B/C boundary / cum. %		
E/N boundary / cum. %		

SUBJECT AGGREGATE	Reference Year	Statistical Equivalence	Recommended Boundaries
Mean/S.D.			
Total Scaled Max. Marks			
A/B boundary / cum. %			
B/C boundary / cum. %			
E/N boundary / cum. %			

APPENDIX 6.1

GWBASIC PROGRAM USED TO ENCODE OBSERVATIONS DURING PHASES 1 AND 2

```
10 CLS
20 PRINT "IMMM
```

```
410 PRINT"                    IMMMMMMMMMMMMMMMMMMMM:"
420 PRINT"                    :   1   Speaks (Neut)  :"
430 PRINT"                    :   2   Agrees         :"
440 PRINT"                    :   3   Disagrees      :"
450 PRINT"                    :   4   Laughs         :"
460 PRINT"                    :   5   Enters room    :"
470 PRINT"                    :   6   Leaves room    :"
480 PRINT"                    :   7   Takes script   :"
490 PRINT"                    :  99   DELETE Entry   :"
500 PRINT"                    HMMMMMMMMMMMMMMMMMMMM<"
510 PRINT
520 INPUT "Action";AC
530 IF AC=99 GOTO 210
540 IF AC>3 AND AC<8 GOTO 220
550 IF AC<1 OR AC>7 GOTO 510
560 GOSUB 580
570 GOTO 220
580 CLS:PRINT:PRINT TIM$,PL,AC,C$:PRINT:PRINT
590 PRINT"       IMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMM:"
600 PRINT"       :   +1   +tive AFFECT/SOCIAL    -1   -tive AFFECT/SOCIAL   :"
610 PRINT"       :   +2   Gives METHOD. GUIDE    -2   Seeks METHOD. GUIDE   :"
620 PRINT"       :   +3   Gives EVAL. CRITERION  -3   Seeks EVAL. CRITERION :"
630 PRINT"       :   +4   Gives PROCEED SUGGEST  -4   Seeks PROCEED SUGGEST :"
640 PRINT"       :   +5   Gives OVERALL JUDGE.   -5   Seeks OVERALL JUDGE.  :"
650 PRINT"       :   +6   Gives EVAL. OF SCRIPT  -6   Seeks EVAL. OF SCRIPT :"
660 PRINT"       :   +7   Gives STAT. OPIN/INFO  -7   Seeks STAT. OPIN/INFO :"
670 PRINT"       :   +8   Gives PAPER/TASK INFO  -8   Seeks PAPER/TASK INFO :"
680 PRINT"       :   +9   Makes OTHER REL. POINT -9   Seeks OTHER REL. POINT:"
690 PRINT"       :   +0   Suggests BOUNDARY      -0   Asks for BOUNDARY     :"
700 PRINT"       HMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMM<"
710 PRINT
720 INPUT "Category";C$
730 BEG=1
740 BTEST=INSTR(BEG,C$,"0")
750 IF BTEST=0 THEN RETURN
760 IF BTEST=1 GOTO 1000
770 IF BTEST>1 AND MID$(C$,BTEST-1,1)<>"-" GOTO 1000
780 BEG=BTEST+1
790 GOTO 740
800 RETURN
810 CLS:PRINT:PRINT:
820 PRINT"                    IMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMM:"
830 PRINT"                    :  END OF RUN - RECORDING FINISHED  :"
840 PRINT"                    HMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMM<"
850 PRINT:PRINT:
860 END
1000 PRINT:PRINT TIM$,PL,AC,C$:PRINT
1005 INPUT"value of suggested boundary":B
1010 GOTO 800
1050 INPUT"acknowledge";XX
```

APPENDIX 6.2
QUANTITATIVE DATA FROM PHASE 2 OBSERVATIONS

Direction	Nature of remark	Accounting	English	General studies	History	Maths	Physics	Commun. studies	Total
Seeks	Other relevant point	10	2	3	5	1	5	2	**28**
	Paper/task opin/info	17	6	12	6	3	6	2	**52**
	Statistical opin/info	17	0	3	21	35	18	9	**103**
	Script evaluation	19	9	5	5	11	0	5	**54**
	Overall judgement	7	3	6	3	5	4	3	**31**
	Procedural suggestion	20	7	9	7	9	6	3	**61**
	Evaluative criterion	4	0	0	0	0	0	0	**4**
	Methodological guidance	4	0	4	3	3	3	1	**18**
	Affective	0	0	4	0	0	1	0	**5**
	Boundary suggestion	13	7	19	17	15	15	7	**93**
Gives	Boundary suggestion	46	25	96	33	57	60	16	**333**
	Affective	53	52	273	40	59	90	52	**619**
	Methodological guidance	30	18	21	20	38	16	18	**161**
	Evaluative criterion	39	4	13	4	4	8	3	**75**
	Procedural suggestion	65	55	32	31	37	35	39	**294**
	Overall judgement	38	24	45	16	24	25	13	**185**
	Script evaluation	167	74	74	41	26	8	53	**443**
	Statistical opin/info	47	8	48	49	153	58	42	**405**
	Paper/task opin/info	80	27	35	18	50	40	10	**260**
	Other relevant point	38	25	16	8	49	43	16	**195**
	Total	**714**	**346**	**718**	**327**	**579**	**441**	**294**	**3419**

APPENDIX 6.3

FORM USED BY AWARDERS TO RECORD THEIR JUDGEMENTS DURING PHASE 2

GRADING PROCESSES PROJECT

Grading Proforma

Subject ..... **Awarder** .....

Please give the following information for each script you study.

Paper Number	Centre Number	Candidiate Number	Your grade judgement	Reasons/Comments

APPENDIX 6.4

QUANTITATIVE DATA ON AWARDERS' EVALUATIVE REASONS - PHASE 2

Reasons		Physics	Maths	History	General studies	Commun. Studies	Accounting	English	Total
Holistic	Moral & Social reasons	0	0	3	2	2	0	2	**7**
	Affective "reasons"	14	47	16	25	24	61	31	**218**
	Genetic reasons	2	2	11	12	14	3	11	**55**
	Stylistic Unity, Structure & Complexity	0	0	9	1	7	0	10	**27**
	Script balance	26	37	4	8	5	25	4	**109**
	Marks	4	5	1	1	0	10	0	**21**
	Script content	8	32	19	3	16	2	13	**93**
Fragmented	Moral & Social reasons	0	0	1	1	1	0	0	**3**
	Affective "reasons"	10	15	5	9	4	5	8	**56**
	Genetic reasons	0	0	0	0	0	0	0	**0**
	Stylistic Unity, Structure & Complexity	0	0	0	0	1	0	3	**4**
	Script balance	2	10	0	0	0	0	0	**12**
	Marks	1	2	2	0	0	6	0	**11**
	Script content	5	9	0	2	2	2	3	**23**
Total		**72**	**159**	**70**	**63**	**76**	**114**	**85**	**639**

APPENDIX 6.5

OUTCOMES DATA FOR 1989, 1990, 1991, 1993 AND 1994

1989

Subject	No. of candidates	Cumulative % at Grade A	Cumulative % at Grade B	Cumulative % at Grade E
Acounting (600) - All	6866	3.6	16.9	56.7
Applied Mathematics (602) - All	1501	30.3	48.1	80.9
Biology (607) - All	3363	11.1	27.7	84.4
Biology (677) - All	1495	10.3	26.1	78.0
Business Studies (655) - All	8947	1.8	8.6	62.8
Chemistry (654) - All	3510	7.5	23.0	70.2
Comm Studies (608) - All	3647	1.0	9.4	78.1
Computing (643) - All	3597	2.5	12.1	73.7
Constitutional Law (612) - All	857	0.5	1.2	25.1
Economic & Soc Hist (620) - All	1113	1.2	10.5	65.9
Economics (618) - All	12598	5.4	13.6	59.2
English I (Lang & Lit) (623) - All	9874	1.9	13.5	81.1
English II (Literature) (652) - All	5303	6.2	16.0	72.7
English III (Lit. Alt.) (660) - All	6702	8.8	26.2	97.7
Environ Science (657) - All	479	6.7	20.3	84.6
French (624) - All	3077	10.8	29.0	78.4
General Studies (667) - All	2546	0.9	7.7	69.2
Geography (626) - All	2519	5.4	19.7	76.2
German (629) - All	1245	15.1	39.0	85.6
Government & Politics (609) - All	1582	5.4	26.3	81.3
History (630) - All	4340	5.4	19.0	78.1
History (Alternative) (673) - All	1637	7.0	23.2	84.7
History of Art (606) - All	1357	2.7	12.3	66.5
Human Biology (642) - All	2716	5.7	16.4	66.2
Law (625) - All	3637	4.0	11.4	49.9
Philosophy (661) - All	446	9.2	24.0	73.3
Photography (634) - All	1136	5.0	18.6	78.2
Physics (635) - All	8140	10.9	23.7	64.2
Psychology (651) - All	7047	5.4	21.1	78.2
Pure & Applied Maths (636) - All	6880	11.6	24.9	67.8
Pure Mathematics (632) - All	3769	21.0	38.1	72.9
Pure Maths & Stats (646) - All	5663	4.3	15.2	60.0
Sociology (639) - All	18741	5.5	20.4	63.3
Spanish (640) - All	552	25.2	52.2	92.9
Sport Studies (665) - All	233	1.7	11.6	79.4
Statistics (641) - All	2364	9.7	24.4	67.6
Theatre Studies (653) - All	4287	3.2	17.1	85.7
Physical Education (703) - All	150	2.7	7.4	91.3

1990

Subject	Max mark	Mean mark	SD of marks	Grade A/B Boundary	Grade B/C boundary	Grade E/N boundary	No. of candidates	Cumulative % at Grade A	Cumulative % at Grade B	Cumulative % at Grade E
Acounting (600) - All	400	176.4	63.5	275	237	167	6775	4.5	18.5	58.6
Schools							692	4.5	23.6	70.3
FE							3697	4.8	20.3	62.0
Overseas							2066	3.7	13.1	48.3
Male							3180	5.7	20.3	59.7
Female							3595	3.4	16.8	57.3
Applied Mathematics (602) - All	210	131.0	50.0	159	138	84	1359	35.8	50.1	81.2
Schools							614	47.2	64.0	91.3
FE							292	39.7	51.0	82.9
Overseas							402	16.7	29.9	66.4
Male							1087	34.7	48.5	80.1
Female							272	40.1	56.3	85.4
Biology (607) - All	260	142.9	31.1	186	165	115	4152	9.2	25.7	80.9
Schools							2542	10.7	27.9	81.1
FE							1568	6.9	22.2	80.6
Overseas							0	0.0	0.0	0.0
Male							1604	10.3	27.9	82.5
Female							2548	8.4	24.1	79.7
Biology (677) - All	260	130.8	35.9	178	157	105	2195	10.3	25.8	76.7
Schools							1040	14.5	33.2	80.6
FE							576	5.6	17.9	73.0
Overseas							397	9.8	23.2	77.0
Male							908	11.8	28.3	78.7
Female							1287	9.2	24.0	75.3

1990

Business Studies (655) - All	300	126.3	35.6	192	169	116	11206	2.6	11.4	64.0
Schools							5235	3.8	15.7	77.9
FE							4109	2.0	9.6	61.6
Overseas							1383	0.1	0.7	18.2
Male							5337	3.3	12.7	67.4
Female							5869	2.0	10.3	60.9
Chemistry (654) - All	400	215.4	58.7	295	264	174	3902	9.5	22.7	74.1
Schools							1596	11.2	24.9	75.7
FE							1779	8.3	21.6	73.6
Overseas							434	10.1	23.7	75.6
Male							2359	10.2	24.3	74.9
Female							1543	8.4	20.1	72.7
Comm Studies (608) - All	500	238.0	44.9	335	288	203	3945	1.5	13.1	79.1
Schools							1274	1.4	13.6	82.1
FE							2654	1.5	12.8	77.9
Overseas							1	0.0	0.0	0.0
Male							1369	1.2	11.2	74.3
Female							2576	1.6	14.1	81.8
Computing (643) - All	500	233.6	68.6	343	301	195	3445	5.6	17.5	71.3
Schools							1498	6.4	19.0	76.2
FE							1865	5.1	16.6	67.9
Overseas							12	0.0	33.3	75.0
Male							2835	6.2	19.4	73.7
Female							610	2.5	8.6	60.1
Constitutional Law (612) - All	200	56.0	24.4	118	105	75	701	0.1	2.0	24.5
Schools							4	0.0	0.0	75.0
FE							25	0.0	4.0	48.0
Overseas							619	0.2	2.0	23.5
Male							433	0.2	2.3	25.6
Female							268	0.0	1.5	22.7

1990

Economic & Soc Hist (620) - All	200	92.0	22.6	140	123	83	928	2.2	8.6	67.9
Schools							129	3.9	14.0	80.7
FE							784	1.9	7.5	65.6
Overseas							0	0.0	0.0	0.0
Male							338	2.7	10.1	66.3
Female							590	1.9	7.7	68.8
Economics (618) - All	200	84.4	28.8	128	111	72	12056	7.8	19.3	64.7
Schools							5488	10.1	24.6	74.3
FE							4195	6.5	16.8	64.0
Overseas							2007	4.1	10.5	43.2
Male							6906	8.2	21.0	67.7
Female							5150	7.2	17.0	61.6
English I (Lang & Lit) (623) - All	300	163.2	35.7	222	197	127	11196	5.1	17.7	85.7
Schools							3147	6.3	20.7	91.0
FE							7782	4.5	16.5	83.9
Overseas							30	13.3	20.0	70.0
Male							2994	6.7	19.8	83.5
Female							8202	4.5	16.9	86.3
English II (Literature) (652) - All	300	128.2	41.6	198	170	103	4401	5.5	16.2	71.8
Schools							1601	6.2	18.5	76.2
FE							2623	5.0	14.6	69.2
Overseas							57	8.8	19.3	54.4
Male							1193	6.4	17.4	68.7
Female							3208	5.2	15.8	73.0
English III (Lit. Alt.) (660) - All	360	224.3	39.4	288	253	151	10061	6.3	24.0	97.6
Schools							7651	6.7	24.9	97.7
FE							2313	4.8	20.6	96.9
Overseas							10	20.0	50.0	100.0
Male							2550	7.7	26.1	97.0
Female							7511	5.8	23.3	97.8

1990

Environ Science (657) - All	250	117.8	26.4	163	142	95	562	4.1	19.4	82.0
Schools							211	3.3	13.7	78.6
FE							326	4.3	22.4	83.5
Overseas							0	0.0	0.0	0.0
Male							279	4.3	21.1	83.1
Female							283	3.9	17.7	80.9
French (624) - All	500	248.4	66.7	332	288	196	4492	11.5	28.3	77.6
Schools							3007	12.3	29.6	78.9
FE							1404	9.2	24.9	74.7
Overseas							16	43.8	68.8	93.9
Male							1022	15.9	34.0	78.6
Female							3470	10.2	26.6	77.3
General Studies (6667) - All	400	196.7	40.4	277	248	170	1141	1.8	11.2	76.6
Schools							849	2.1	11.2	76.9
FE							280	0.7	11.1	75.0
Overseas							9	11.1	22.2	77.7
Male							510	2.4	12.0	76.9
Female							631	1.4	10.6	76.2
Geography (626) - All	600	260.2	58.7	358	321	230	2813	5.4	15.4	69.8
Schools							1969	5.6	16.5	70.8
FE							795	5.2	13.0	67.6
Overseas							23	4.3	17.3	78.1
Male							1579	5.2	14.6	69.4
Female							1234	5.7	16.4	70.4
German (629) - All	500	277.2	66.9	350	297	202	1706	14.8	37.7	87.1
Schools							1222	15.0	39.0	88.8
FE							452	13.1	33.0	82.3
Overseas							13	38.5	61.6	100.1
Male							443	20.8	45.2	86.9
Female							1263	12.7	35.1	87.3

1990

Government & Politics (609) - All	200	119.5	29.0	156	137	96	1658	8.6	29.8	81.8
Schools							704	9.4	34.7	87.4
FE							847	9.0	28.4	80.3
Overseas							23	4.3	4.3	30.3
Male							967	9.0	30.6	81.4
Female							691	8.1	28.6	82.0
History (630) - All	300	95.4	25.6	136	117	76	5234	5.3	20.7	79.0
Schools							2778	7.2	25.2	82.7
FE							2212	3.2	15.6	74.9
Overseas							88	1.1	12.5	73.9
Male							2348	6.2	23.0	80.7
Female							2886	4.5	18.7	77.7
History (Alternative) (673) - All	400	200.2	45.1	271	235	151	2125	6.4	22.2	86.8
Schools							1820	6.8	23.6	88.0
FE							241	3.7	13.2	78.3
Overseas							0	0.0	0.0	0.0
Male							862	6.6	24.1	86.2
Female							1263	6.3	20.9	87.3
History of Art (606) - All	200	86.8	26.8	139	118	68	1233	2.3	12.8	75.7
Schools							420	1.4	14.5	80.5
FE							661	2.1	9.2	71.0
Overseas							19	10.5	57.9	73.7
Male							325	1.8	9.2	69.2
Female							908	2.4	14.0	78.0
Human Biology (642) - All	390	163.6	56.1	264	232	147	2832	5.2	12.9	59.0
Schools							470	4.0	10.8	57.6
FE							2316	5.4	13.3	59.1
Overseas							0	0.0	0.0	0.0
Male							600	7.7	15.7	62.0
Female							2232	4.5	12.1	58.2

1990

Law (625) - All	400	150.5	53.3	215	195	135	3747	11.2	20.9	62.2
Schools							271	15.1	28.0	74.9
FE							2369	13.5	24.4	68.2
Overseas							914	4.6	10.3	42.3
Male							1646	9.9	18.3	59.6
Female							2101	12.1	22.8	64.2
Music (633) - All	800	415.5	96.6	578	492	352	414	8.0	21.3	74.9
Schools							240	7.1	22.5	76.3
FE							167	9.0	18.6	73.2
Overseas							0	0.0	0.0	0.0
Male							130	13.1	28.5	72.3
Female							284	5.6	17.9	76.0
Philosophy (661) - All	200	87.8	30.8	132	114	69	476	8.4	19.1	72.8
Schools							141	12.8	29.8	85.8
FE							313	6.7	14.7	67.8
Overseas							1	0.0	0.0	100.0
Male							258	7.8	17.9	72.2
Female							218	9.2	20.7	73.9
Photography (634) - All	300	152.9	34.7	208	180	120	1335	5.8	23.0	82.8
Schools							277	1.8	13.7	74.7
FE							974	7.0	26.1	85.6
Overseas							25	0.0	8.0	60.0
Male							816	7.4	24.6	84.2
Female							519	3.3	20.6	80.7
Physics A (635) - All	440	214.8	62.9	294	259	185	7412	11.5	25.2	66.3
Schools							3310	14.3	29.7	70.2
FE							2788	9.5	21.8	63.4
Overseas							1107	9.0	23.1	65.7
Male							5885	12.2	26.3	66.5
Female							1527	8.9	21.4	65.9

1990

Psychology (651) - All	500	243.1	68.2	354	304	200	8504	4.3	19.8	73.7
Schools							1506	5.1	22.0	80.2
FE							6916	4.1	19.4	72.4
Overseas							0	0.0	0.0	0.0
Male							1924	4.2	17.6	68.6
Female							6580	4.3	20.5	75.2
Pure & Applied Maths (636) - All	210	105.3	45.9	161	141	85	6718	12.6	25.3	66.5
Schools							3626	14.6	27.8	69.5
FE							2742	9.4	20.9	61.8
Overseas							207	24.6	45.4	86.5
Male							5040	12.1	24.1	65.1
Female							1678	14.1	28.8	70.3
Pure Mathematics (632) - All	210	126.0	52.9	160	137	84	3219	31.7	48.5	77.1
Schools							1034	48.0	64.4	87.9
FE							1177	23.8	40.5	71.9
Overseas							876	26.4	44.3	75.4
Male							2209	32.2	48.9	77.6
Female							1010	30.7	47.6	76.5
Pure Maths & Stats (646) - All	207	93.1	37.6	155	137	80	5464	4.3	13.2	65.0
Schools							2635	4.9	14.4	67.9
FE							2325	3.2	10.9	58.8
Overseas							353	7.4	20.1	84.4
Male							2735	3.9	11.9	61.8
Female							2729	4.8	14.5	68.1
Sociology (639) - All	200	82.3	25.9	122	106	76	19789	5.2	19.1	62.0
Schools							7955	5.8	21.7	66.6
FE							11043	4.9	17.9	60.0
Overseas							280	0.7	1.1	22.8
Male							5126	5.7	20.2	60.6
Female							14663	5.0	18.7	62.4

1990

Spanish (640) - All	500	297.0	68.2	353	304	208	773	22.9	50.1	90.2
Schools							420	24.5	50.2	92.4
FE							323	18.6	47.4	86.7
Overseas							2	50.0	100.0	100.0
Male							198	34.3	60.1	90.9
Female							575	19.0	46.7	90.0
Sport Studies (665) - All	250	105.6	22.6	163	136	83	477	0.6	7.3	85.1
Schools							164	1.2	10.3	91.4
FE							310	0.3	5.8	81.7
Overseas							0	0.0	0.0	0.0
Male							276	1.1	8.7	86.3
Female							201	0.0	5.5	83.6
Statistics (641) - All	204	83.9	40.7	146	129	72	1939	7.7	15.6	59.6
Schools							499	10.8	20.0	65.8
FE							675	3.7	8.7	55.6
Overseas							649	10.6	20.9	60.7
Male							1062	7.9	16.5	59.3
Female							877	7.4	14.6	60.3
Theatre Studies (653) - All	250	129.0	28.5	182	159	104	4749	3.4	15.3	81.6
Schools							2863	3.6	16.6	82.8
FE							1804	2.8	12.9	79.5
Overseas							0	0.0	0.0	0.0
Male							1160	3.5	14.7	76.6
Female							3589	3.3	15.4	83.0
Physical Education (703) - All	333	184.5	26.0	242	218	152	222	0.9	9.9	90.9
Schools							86	2.3	12.8	90.8
FE							136	0.0	8.1	91.2
Overseas							0	0.0	0.0	0.0
Male							124	0.0	7.3	94.4
Female							98	2.0	13.2	86.6

1991

Subject	Max mark	Mean mark	SD of marks	Grade A/B Boundary	Grade B/C boundary	Grade E/N boundary	No. of candidates	Cumulative % at Grade A	Cumulative % at Grade B	Cumulative % at Grade E
Acounting (600) - All	400	156.6	59.7	266	223	144	6637	3.1	14.6	58.5
Schools							804	3.5	17.2	71.8
FE							3663	3.3	16.7	62.8
Overseas							1835	2.1	8.4	43.6
Male							3135	3.4	15.5	59.3
Female							3502	2.7	13.6	57.5
Applied Mathematics (602) - All	210	125.5	47	158	136	82	1237	29.3	46.7	81.3
Schools							515	43.3	62.5	95.5
FE							276	30.8	52.9	81.9
Overseas							405	10.9	22.8	63.3
Male							994	28.9	46.2	81.1
Female							243	30.9	48.6	82.8
Biology (607) - All	260	133.9	29	185	164	119	5159	5.0	16.3	68.5
Schools							3266	5.5	17.5	70.0
FE							1846	4.3	14.6	66.5
Overseas							0	0.0	0.0	0.0
Male							1887	5.2	16.7	69.4
Female							3272	4.9	16.1	68.1
Biology (677) - All	260	112.9	33.1	168	146	100	2464	5.3	17.4	64.0
Schools							1121	7.3	22.3	72.5
FE							732	3.0	12.3	57.5
Overseas							456	4.6	16.7	58.2
Male							1020	5.3	17.5	65.2
Female							1444	5.3	17.4	63.2

1991

Business Studies (655) - All	300	136.5	36.9	199	175	119	12477	3.9	14.6	70.2
Schools							6080	5.0	17.9	78.8
FE							4801	3.6	14.0	71.0
Overseas							1021	0.0	0.8	21.2
Male							6061	3.6	14.0	70.3
Female							6416	4.2	15.2	70.3
Chemistry (654) - All	400	195.7	60.4	289	258	168	4139	6.8	16.9	66.1
Schools							1472	8.4	18.9	68.6
FE							1977	4.9	15.1	65.9
Overseas							606	10.2	20.4	65.1
Male							2486	7.9	18.6	67.7
Female							1653	5.3	14.7	64.0
Comm Studies (608) - All	500	255	48.1	335	288	203	4565	4.0	19.5	81.2
Schools							1407	4.4	22.2	84.1
FE							3110	3.9	18.5	80.2
Overseas							1	0.0	0.0	100.0
Male							1546	3.0	15.8	75.4
Female							3019	4.6	21.5	84.3
Computing (643) - All	333	150.2	41.7	230	199	126	3294	3.0	13.3	70.6
Schools							1569	3.3	13.8	72.8
FE							1644	2.9	13.0	68.6
Overseas							4	0.0	0.0	100.0
Male							2694	3.5	14.7	72.6
Female							600	1.0	7.2	62.5
Constitutional Law (612) - All	200	64.4	23.9	117	106	74	669	0.1	2.8	38.1
Schools							27	3.7	11.1	51.8
FE							27	0.0	3.7	59.2
Overseas							560	0.0	2.5	37.0
Male							393	0.0	3.6	39.0
Female							276	0.4	1.8	37.0

1991

Economic & Soc Hist (620) - All	200	88.5	22.5	138	122	82	1066	1.0	6.0	63.5
Schools							125	2.4	8.0	64.8
FE							905	0.8	5.7	63.4
Overseas							0	0.0	0.0	0.0
Male							435	1.1	6.8	61.1
Female							631	1.0	5.4	65.2
Economics (618) - All	200	84.2	31.3	130	114	72	11913	7.9	18.5	63.8
Schools							5150	11.2	24.1	73.7
FE							4209	6.4	16.6	62.4
Overseas							2159	3.4	9.7	42.8
Male							6861	8.6	19.7	66.3
Female							5052	7.0	16.9	60.1
English I (Lang & Lit) (623) - All	300	156.4	34.8	216	192	122	12186	4.8	15.6	84.8
Schools							3350	6.1	19.4	90.1
FE							8538	4.3	14.3	83.1
Overseas							36	8.3	16.6	52.8
Male							3318	6.4	17.9	83.5
Female							8868	4.2	14.7	85.4
English II (Literature) (652) - All	300	118.9	46.7	245	204	89	3680	0.7	4.8	71.9
Schools							1150	0.7	6.0	76.5
FE							2312	0.6	4.0	69.9
Overseas							42	2.4	9.5	54.7
Male							1029	1.2	4.9	66.7
Female							2651	0.5	4.8	74.0
English III (Lit. Alt.) (660) - All	360	212.3	47.3	276	245	141	13929	10.5	24.9	94.0
Schools							10621	10.9	25.4	94.4
FE							3137	9.0	22.9	92.6
Overseas							13	7.7	38.5	100.0
Male							3694	11.0	25.3	92.7
Female							10235	10.3	24.7	94.5

1991

Environ Science (657) - All	250	118.9	27.3	167	144	95	794	3.9	19.8	83.4
Schools							341	2.9	18.1	82.1
FE							422	4.7	20.3	84.0
Overseas							0	0.0	0.0	0.0
Male							397	5.3	24.2	87.4
Female							397	2.5	15.3	79.3
French (624) - All	500	266.5	66.9	347	295	202	5324	12.7	33.7	82.7
Schools							3460	13.6	35.5	84.6
FE							1736	11.0	30.2	79.6
Overseas							14	28.6	50.0	78.6
Male							1362	14.5	35.9	84.0
Female							3962	12.1	33.0	82.4
General Studies (6667) - All	400	209.4	39.8	279	246	168	1256	4.6	17.8	85.5
Schools							991	4.8	18.2	86.2
FE							175	4.6	15.5	81.8
Overseas							13	15.4	30.8	100.1
Male							548	5.5	17.0	82.7
Female							708	4.0	18.5	87.7
Geography (626) - All	600	273.4	55.1	355	311	229	3026	6.8	24.2	78.8
Schools							2107	7.5	26.2	80.2
FE							862	5.1	19.4	75.1
Overseas							29	10.3	31.0	96.5
Male							1698	6.9	23.1	79.3
Female							1328	6.7	25.6	78.1
German (629) - All	500	238.8	75.8	331	280	177	2140	13.2	30.7	77.1
Schools							1515	13.9	32.6	78.0
FE							577	10.7	25.6	75.0
Overseas							9	33.3	55.5	77.7
Male							610	17.0	37.0	80.6
Female							1530	11.6	28.1	75.8

1991

Government & Politics (609) - All	200	118.8	27.8	155	137	94	1835	8.7	26.9	83.6
Schools							739	10.7	31.5	90.9
FE							976	7.7	25.1	79.7
Overseas							14	0.0	0.0	35.7
Male							1060	9.2	27.6	83.4
Female							775	8.1	26.0	84.1
History (630) - All	200	92.9	26.1	135	117	77	5894	5.3	18.5	74.8
Schools							3065	7.1	23.3	81.5
FE							2550	3.3	13.3	67.4
Overseas							92	5.4	10.8	63.0
Male							2571	6.2	20.9	76.0
Female							3323	4.5	16.4	73.8
History (Alternative) (673) - All	400	194.5	46.1	270	234	151	2397	5.8	20.6	83.4
Schools							2132	6.1	20.9	84.2
FE							220	3.6	20.0	77.7
Overseas							0	0.0	0.0	0.0
Male							940	4.9	19.8	84.3
Female							1457	6.4	21.2	82.9
History of Art (606) - All	200	92.4	27.1	138	116	67	1226	4.2	21.3	81.7
Schools							438	4.3	22.3	85.1
FE							661	3.8	19.2	78.2
Overseas							13	30.8	84.6	100.0
Male							329	3.3	17.6	76.0
Female							897	4.5	22.7	83.8
Human Biology (642) - All	390	165.2	54.4	262	232	148	3300	5.2	13.0	59.4
Schools							671	4.2	11.9	61.2
FE							2555	5.5	13.3	59.2
Overseas							0	0.0	0.0	0.0
Male							692	8.2	18.0	63.6
Female							2608	4.3	11.6	58.0

1991

Law (625) - All	200	77.6	25.3	114	102	71	4166	7.0	17.1	62.1
Schools							390	9.2	22.5	75.6
FE							2615	8.0	18.9	66.0
Overseas							982	4.5	11.4	48.9
Male							1776	6.5	16.2	58.3
Female							2390	7.4	17.8	64.9
Music (633) - All	800	410.2	98.5	563	484	341	394	6.6	22.8	73.8
Schools							196	4.6	25.0	74.0
FE							180	8.3	20.0	75.1
Overseas							0	0.0	0.0	0.0
Male							139	9.4	25.9	74.1
Female							255	5.1	21.2	73.8
Philosophy (661) - All	200	79.6	31.8	139	115	74	699	4.6	15.0	57.5
Schools							160	5.0	20.0	67.4
FE							441	3.9	12.5	53.1
Overseas							2	0.0	0.0	100.0
Male							339	5.3	15.3	57.2
Female							360	3.9	14.7	57.8
Photography (634) - All	300	152.6	34.2	206	178	121	1307	6.7	23.4	82.8
Schools							271	4.1	17.4	79.4
FE							975	7.6	25.3	83.9
Overseas							9	0.0	0.0	33.3
Male							771	7.8	26.9	85.5
Female							536	5.2	18.4	79.0
Physics A (635) - All	440	208.5	68.4	298	263	184	7423	11.5	22.4	61.1
Schools							3111	14.3	25.3	62.7
FE							2833	9.8	20.4	58.8
Overseas							1305	9.1	20.4	62.9
Male							5764	12.3	23.2	61.4
Female							1659	9.0	20.0	60.7

1991

Psychology (651) - All	500	234	70	343	292	179	10120	5.4	21.0	77.7
Schools							1923	6.8	23.6	83.7
FE							8009	5.2	20.4	76.4
Overseas							9	0.0	0.0	11.1
Male							2211	5.2	19.2	72.4
Female							7909	5.5	21.6	79.4
Pure & Applied Maths (636) - All	210	102.4	46.7	160	137	82	6647	12.3	25.9	66.3
Schools							3377	13.9	27.4	69.6
FE							2876	8.9	22.6	61.4
Overseas							241	34.9	53.6	88.0
Male							4922	12.1	25.3	65.6
Female							1725	12.8	27.6	68.1
Pure Mathematics (632) - All	210	96.5	46.8	144	122	65	3137	18.5	32.8	72.7
Schools							935	28.9	47.6	83.9
FE							991	14.3	28.6	67.2
Overseas							1095	14.3	26.1	70.0
Male							2142	18.5	32.9	72.7
Female							995	18.4	32.6	72.7
Pure Maths & Stats (646) - All	207	90.4	37.3	153	131	78	5432	4.7	14.8	63.6
Schools							2388	4.9	15.7	66.3
FE							2373	3.8	12.3	57.3
Overseas							549	7.8	22.6	81.8
Male							2669	4.5	13.9	60.6
Female							2763	5.0	15.7	66.4
Sociology (639) - All	200	85.2	24.7	124	110	78	17222	5.7	16.4	62.7
Schools							6543	7.1	19.5	69.8
FE							9884	5.0	14.9	59.4
Overseas							212	0.0	1.4	9.9
Male							4662	5.9	15.8	59.0
Female							12560	5.6	16.6	64.2

1991

Spanish (640) - All	500	270.4	74.6	344	297	206	928	18.9	38.3	79.5
Schools							538	18.4	38.3	81.1
FE							360	18.0	36.6	76.0
Overseas							5	100.0	100.0	100.0
Male							256	28.5	48.8	86.3
Female							672	15.2	34.2	76.9
Sport Studies (665) - All	250	105.9	23.4	166	143	80	749	0.3	5.5	87.5
Schools							190	0.0	2.6	91.6
FE							550	0.4	6.0	85.8
Overseas							0	0.0	0.0	0.0
Male							404	0.2	4.9	88.1
Female							345	0.3	6.1	86.6
Statistics (641) - All	204	100.8	40.4	152	131	80	1778	9.7	25.4	69.5
Schools							421	13.5	29.9	76.0
FE							608	7.9	23.5	69.4
Overseas							651	8.9	24.7	65.8
Male							956	10.3	27.4	69.3
Female							822	9.1	23.2	69.9
Theatre Studies (653) - All	250	129	28.7	185	160	105	5634	2.5	14.8	80.2
Schools							3500	2.8	16.2	82.9
FE							1991	1.9	12.3	75.8
Overseas							0	0.0	0.0	0.0
Male							1456	2.0	14.6	74.3
Female							4178	2.7	14.9	82.3
Physical Education (703) - All	333	172.9	31	240	210	135	630	1.3	11.5	89.8
Schools							351	1.4	12.2	91.2
FE							258	0.4	9.3	88.8
Overseas							0	0.0	0.0	0.0
Male							339	0.6	10.3	91.4
Female							291	2.1	12.8	88.0

1993

Subject	Max mark	Mean mark	SD of marks	Grade A/B Boundary	Grade B/C boundary	Grade E/N boundary	No. of candidates	Cumulative % at Grade A	Cumulative % at Grade B	Cumulative % at Grade E
Acounting (600) - All	400	175.5	57.9	253	221	161	4630	8.5	24.2	61.8
Schools							460	6.5	25.6	66.4
FE							3496	9.2	24.9	62.9
Overseas							108	0.0	7.4	50.9
Male							2316	8.5	24.3	62.1
Female							2314	8.6	24.2	61.6
Applied Mathematics (602) - All	210	119.4	51.0	144	119	66	836	35.8	54.3	83.9
Schools							344	50.3	68.3	93.3
FE							252	34.1	56.3	86.0
Overseas							196	15.8	31.6	67.4
Male							647	35.4	53.8	83.2
Female							189	37.0	56.0	86.2
Biology (607) - All	260	126.6	32.4	169	144	99	6333	10.9	30.4	79.9
Schools							3153	11.9	32.2	82.3
FE							2750	10.0	28.4	77.1
Overseas							3	0.0	33.3	66.6
Male							2563	11.9	31.8	82.3
Female							3770	10.2	29.3	78.4
Biology (677) - All	260	111.4	38.1	158	138	84	2350	12.7	25.9	74.7
Schools							699	17.5	30.5	80.3
FE							1026	9.3	21.5	71.7
Overseas							239	18.0	32.2	70.7
Male							942	15.0	28.9	78.3
Female							1408	11.2	24.0	72.3

1993

Business Studies (655) - All	300	121.1	31.0	171	149	102	14863	5.6	19.1	74.8
Schools							4722	6.2	20.8	78.9
FE							8403	5.3	18.2	72.9
Overseas							113	6.2	19.5	57.5
Male							7611	5.6	18.7	74.9
Female							7252	5.6	19.1	74.8
Chemistry (0654A) All	400	201.7	59.1	283	253	167	1924	9.5	22.5	70.6
Schools							738	10.7	24.9	74.1
FE							900	9.4	22.1	69.7
Overseas							113	6.2	15.9	54.0
Male							1170	9.8	22.9	70.6
Female							754	9.0	22.0	70.9
Chemistry (0654B) - All	400	211.4	61.5	288	258	172	1552	11.9	25.3	72.3
Schools							300	12.0	26.3	74.2
FE							754	8.1	20.4	64.8
Overseas							377	19.1	35.3	83.3
Male							896	13.7	27.1	73.6
Female							656	9.3	22.7	70.1
Comm Studies (608) - All	500	243.3	50.8	301	267	199	5303	12.9	32.0	82.5
Schools							933	14.0	37.5	86.6
FE							4083	12.8	30.9	81.7
Overseas							2	0.0	0.0	100.0
Male							1882	9.6	28.0	76.8
Female							3421	14.7	34.2	85.4
Computing (643) - All	500	214.0	68.8	317	277	171	3988	8.0	19.5	72.4
Schools							1105	9.1	21.1	76.3
FE							2576	7.3	18.7	71.1
Overseas							24	8.3	29.1	79.0
Male							3282	8.7	20.6	74.2
Female							706	5.0	14.9	64.9

1993

Constitutional Law (612) - All	200	61.3	27.4	108	96	69	558	3.8	11.0	41.6
Schools							3	0.0	0.0	66.6
FE							22	4.5	22.7	63.6
Overseas							521	3.8	10.5	40.5
Male							282	2.5	8.5	37.2
Female							276	5.1	13.4	46.0
Economic & Soc Hist (620) - All	200	84.8	21.7	125	113	79	1135	3.3	8.3	64.3
Schools							88	5.7	13.7	72.8
FE							918	2.7	7.8	63.6
Overseas							2	0.0	0.0	0.0
Male							430	4.4	10.2	66.8
Female							705	2.6	7.1	62.8
Economics (618) - All	200	87.6	27.8	129	110	72	9134	7.9	22.3	69.9
Schools							3913	8.8	24.3	72.5
FE							4274	7.3	21.2	68.4
Overseas							123	1.6	7.3	29.3
Male							5868	8.5	23.3	70.3
Female							3266	6.9	20.4	69.1
English I (Lang & Lit) (623) - All	300	156.1	35.9	205	182	123	14540	9.3	24.1	83.2
Schools							2212	11.1	26.8	88.6
FE							10564	9.1	23.9	82.4
Overseas							57	1.8	8.8	61.5
Male							4099	9.1	22.8	80.4
Female							10441	9.3	24.5	84.1
English II (Literature) (652) - All	300	109.3	46.0	171	146	73	2906	10.5	22.1	77.6
Schools							652	12.1	24.1	85.3
FE							1713	9.5	20.7	74.9
Overseas							76	14.5	21.1	64.5
Male							887	12.0	22.4	73.8
Female							2019	9.9	22.0	79.4

1993

English III (Lit. Alt.) (660) - All	360	209.8	44.5	260	233	145	19341	14.6	30.4	94.5
Schools							10322	14.7	29.4	93.8
FE							7537	14.7	31.8	95.5
Overseas							24	25.0	41.7	95.8
Male							5460	14.7	30.1	93.2
Female							13881	14.6	30.5	94.9
Environ Science (657) - All	250	105.5	27.3	148	127	83	1463	6.8	23.4	81.3
Schools							485	4.9	22.0	83.0
FE							839	6.8	23.1	79.9
Overseas							4	25.0	25.0	75.0
Male							735	7.1	25.2	85.6
Female							728	6.5	21.6	77.1
French (624) - All	500	234.1	64.7	297	258	175	5731	16.9	35.8	81.6
Schools							2427	18.5	38.8	84.0
FE							2914	16.0	33.3	80.3
Overseas							34	41.2	82.4	100.0
Male							1497	21.4	40.4	84.6
Female							4234	15.4	34.2	80.6
General Studies (6667) - All	400	213.7	46.0	272	243	164	1901	10.9	27.4	87.3
Schools							1109	9.7	25.1	85.2
FE							602	11.1	27.7	89.9
Overseas							18	22.2	55.5	100.0
Male							837	10.3	25.8	84.2
Female							1064	11.4	28.6	89.7
Geography (626) - All	600	262.8	56.2	323	289	219	3616	14.1	32.8	79.7
Schools							1780	16.3	36.2	83.4
FE							1457	11.4	27.9	74.1
Overseas							37	8.1	40.5	86.4
Male							1986	14.3	32.9	80.6
Female							1630	13.8	32.5	78.3

1993

German (629) - All	500	245.4	74.3	318	274	176	2291	17.8	36.6	81.1
Schools							1120	18.4	38.2	82.4
FE							1038	16.5	34.0	79.0
Overseas							13	53.8	69.2	92.3
Male							649	19.7	40.3	84.0
Female							1642	17.0	35.1	80.0
Government Politics (0609B) All	320	184.2	49.5	242	213	140	1089	10.8	31.5	81.4
Schools							482	11.6	34.4	85.5
FE							497	11.1	30.2	79.5
Overseas							8	12.5	37.5	75.0
Male							669	11.4	33.8	83.7
Female							420	10.0	27.9	77.8
History (0630BM) All	200	94.1	25.2	134	115	78	2433	6.2	21.2	76.6
Schools							855	8.8	26.5	82.8
FE							1281	5.1	19.4	72.4
Overseas							8	0.0	0.0	50.0
Male							1198	7.3	19.9	77.1
Female							1235	5.2	21.2	76.6
History (Alternative) (0673B) All	400	194.2	47.9	258	227	150	1497	9.8	24.6	83.6
Schools							1059	10.5	24.9	83.6
FE							388	8.0	24.5	84.0
Overseas							0	0.0	0.0	0.0
Male							644	10.4	24.2	85.2
Female							853	9.4	25.0	82.5
History of Art (606F) All	200	78.5	25.4	126	107	59	830	3.3	15.7	78.7
Schools							168	1.8	7.8	79.2
FE							573	3.8	17.6	77.6
Overseas							4	0.0	50.0	75.0
Male							210	1.9	13.8	72.8
Female							620	3.7	24.1	78.7

1993

Human Biology (642) - All	390	173.3	52.4	249	223	156	3624	8.4	18.8	60.8
Schools							231	11.7	31.2	70.7
FE							3076	7.9	17.6	60.2
Overseas							5	0.0	0.0	60.0
Male							924	9.2	20.2	61.9
Female							2700	8.2	18.8	60.8
Law (625) - All	200	77.9	25.3	110	98	70	4421	11.2	22.7	64.5
Schools							324	7.1	20.7	66.0
FE							3349	12.2	23.9	65.4
Overseas							78	2.6	10.3	43.7
Male							1796	10.1	20.8	60.5
Female							2625	11.9	23.9	67.2
Music (633) - All	800	389.1	98.7	494	435	320	304	14.5	32.6	74.3
Schools							137	12.4	31.4	73.7
FE							129	17.8	35.6	78.3
Overseas							2	50.0	50.0	50.0
Male							117	16.2	32.4	73.5
Female							187	13.4	32.7	74.9
Philosophy (661) - All	200	82.0	34.1	131	108	63	920	8.6	25.6	70.4
Schools							145	13.1	36.5	83.4
FE							662	7.3	23.0	67.1
Overseas							9	22.2	33.3	100.0
Male							481	10.0	25.2	70.1
Female							439	7.1	26.0	70.9
Photography (634) - All	300	148.7	32.5	191	166	119	1575	10.8	30.0	83.1
Schools							158	9.5	24.1	79.2
FE							1253	11.2	31.2	83.7
Overseas							1	0.0	0.0	100.0
Male							914	11.5	32.5	83.7
Female							661	9.8	26.4	82.3

1993

Physics A (635) - All	500	287.8	77.2	371	336	242	3850	15.9	29.5	71.6
Schools							1694	17.9	33.2	76.0
FE							1461	13.7	25.7	67.6
Overseas							261	14.2	24.9	67.4
Male							3072	16.3	29.5	71.1
Female							778	14.3	29.3	73.6
Physics B (635) - All	500	281.6	77.4	369	334	247	1886	14.4	26.8	67.7
Schools							485	16.7	28.7	70.5
FE							944	13.2	25.5	68.0
Overseas							277	15.2	25.7	61.5
Male							1488	15.1	26.9	67.5
Female							398	11.6	25.9	67.6
Psychology (651) - All	500	242.1	70.2	333	293	191	15452	10.0	25.0	76.8
Schools							1669	10.7	25.6	79.9
FE							12521	10.1	25.1	76.5
Overseas							24	4.2	8.4	45.9
Male							3695	7.8	20.7	69.9
Female							11757	10.7	26.3	78.9
Pure & Applied Maths (636) - All	210	92.4	46.0	149	124	71	5195	13.1	27.5	66.1
Schools							2248	15.7	32.6	70.2
FE							2211	11.2	23.3	62.4
Overseas							100	10.0	25.0	66.0
Male							3853	12.2	26.0	63.7
Female							1342	15.5	31.7	72.6
Pure Mathematics (632) - All	210	123.0	53.7	168	152	91	1944	23.9	35.9	73.3
Schools							554	39.0	54.0	85.8
FE							681	25.0	38.1	74.4
Overseas							506	9.1	18.2	59.8
Male							1356	24.1	36.4	73.2
Female							588	23.5	34.7	73.1

1993

Pure Maths & Stats (646) - All	210	109.7	44.5	169	148	93	3944	9.0	23.1	65.6
Schools							1427	11.2	26.2	68.8
FE							2115	7.4	21.1	64.1
Overseas							44	20.5	31.9	54.6
Male							1938	7.3	21.5	61.3
Female							2006	10.7	24.7	69.8
Sociology (639) - All	200	85.5	26.1	121	108	77	18087	8.4	20.5	64.4
Schools							4506	9.4	23.0	68.4
FE							11065	8.3	20.3	63.7
Overseas							171	2.9	7.0	30.5
Male							4973	7.7	18.4	59.7
Female							13114	8.7	21.3	66.3
Sociology (664) - All	200	100.9	23.3	131	113	86	7741	10.3	32.0	75.3
Schools							2643	10.4	31.7	75.8
FE							4392	10.4	32.2	75.0
Overseas							2	0.0	0.0	100.0
Male							1774	10.0	28.8	71.0
Female							5967	10.4	32.9	76.6
Spanish (640) - All	500	261.9	68.6	321	288	194	1067	20.8	39.5	83.7
Schools							294	27.2	46.9	86.3
FE							671	16.5	34.5	81.6
Overseas							7	57.1	57.1	100.0
Male							287	23.0	43.9	84.7
Female							780	20.0	37.8	83.2
Sport Studies (665) - All	250	103.3	21.9	143	129	80	1616	3.8	13.5	86.8
Schools							338	3.6	11.6	87.1
FE							1150	4.0	14.7	87.4
Overseas							0	0.0	0.0	0.0
Male							996	2.7	9.8	86.4
Female							620	5.5	19.4	92.7

1993

Statistics (641) - All	210	124.2	44.6	170	150	95	1089	15.5	33.3	75.8
Schools							208	17.8	39.4	86.6
FE							551	12.0	28.7	72.2
Overseas							181	26.5	48.6	86.6
Male							681	17.9	34.1	76.3
Female							408	11.5	32.1	75.0
Theatre Studies (653) - All	250	133.4	29.1	172	154	108	7088	9.8	25.1	81.8
Schools							3433	11.3	27.4	84.1
FE							3240	8.1	22.5	79.4
Overseas							3	0.0	0.0	100.0
Male							1788	9.2	21.7	76.1
Female							5300	10.0	26.3	83.8
Physical Education (703) - All	1000	573.7	96.4	702	654	443	2499	8.6	22.4	91.3
Schools							818	7.7	21.5	91.9
FE							1570	9.3	22.9	91.1
Overseas							0	0.0	0.0	0.0
Male							1537	7.9	20.7	91.6
Female							962	9.9	25.3	90.9

1994

Subject	Max mark	Mean mark	SD of marks	Grade A/B Boundary	Grade B/C boundary	Grade E/N boundary	No. of candidates	Cumulative % at Grade A	Cumulative % at Grade B	Cumulative % at Grade E
Acounting (600) - All	400	153.4	57.6	228	198	140	4105	9.7	24.5	60.2
Schools							430	11.6	32.8	72.1
FE							3098	9.5	23.8	58.7
Overseas							42	14.3	23.8	52.4
Male							2075	9.4	23.9	59.2
Female							2030	9.9	25.1	61.3
Applied Mathematics (602) - All	210	128.8	48.6	162	137	76	773	31.2	49.9	83.8
Schools							272	49.3	68.4	95.6
FE							226	30.5	55.3	86.7
Overseas							221	9.0	21.3	69.7
Male							591	30.8	49.9	83.6
Female							182	32.4	50.0	84.6
Biology (607) - All	260	130.0	32.1	171	147	103	7016	12.0	31.3	79.6
Schools							3530	12.5	32.8	80.8
FE							2750	11.4	29.5	78.1
Overseas							30	16.7	43.3	93.3
Male							2798	11.5	30.1	80.3
Female							4218	12.3	32.2	79.2
Biology (677) - All	260	114.7	37.2	160	141	89	2654	13.1	27.3	75.4
Schools							716	15.8	31.4	79.2
FE							1161	9.5	22.1	72.1
Overseas							255	19.2	34.9	82.0
Male							1088	14.8	28.7	76.6
Female							1566	11.9	26.4	74.6

1994

Business Studies (655) - All	300	136.6	34.8	188	166	115	16018	6.8	20.6	74.7
Schools							5365	8.8	24.4	78.9
FE							8923	5.6	18.4	72.6
Overseas							48	2.1	27.1	81.3
Male							8481	7.6	22.1	76.4
Female							7537	6.0	18.9	72.7
Chemistry (0654A) All	400	211.9	61.7	294	264	180	1836	10.2	22.5	68.5
Schools							666	10.8	23.9	73.6
FE							844	8.5	21.0	64.9
Overseas							124	17.7	29.0	71.8
Male							1068	10.6	22.9	68.6
Female							768	9.8	22.0	68.4
Chemistry (0654B) - All	400	221.4	61.6	295	269	180	1426	13.0	26.1	73.4
Schools							265	13.2	21.5	67.5
FE							643	7.9	20.5	65.6
Overseas							384	24.0	42.4	91.4
Male							819	12.8	26.1	71.6
Female							607	13.3	26.0	75.9
Comm Studies (608) - All	500	249.3	48.9	308	275	205	5571	12.0	30.0	82.9
Schools							1009	16.8	38.0	85.9
FE							4283	10.9	28.1	82.0
Overseas							27	37.0	51.9	88.9
Male							2034	8.6	23.5	76.5
Female							3537	14.0	33.7	86.5
Computing (643) - All	500	210.1	73.0	314	277	175	4063	9.2	19.4	68.0
Schools							1204	9.4	18.9	67.4
FE							2592	9.2	20.1	68.7
Overseas							90	4.4	5.6	50.0
Male							3315	10.1	21.2	69.8
Female							748	5.1	11.4	60.2

1994

Constitutional Law (612) - All	200	66.1	27.0	113	100	74	496	3.6	11.5	44.0
Schools							25	0.0	0.0	28.0
FE							12	25.0	25.0	91.7
Overseas							426	3.3	12.4	44.4
Male							228	2.6	11.0	45.2
Female							268	4.5	11.9	42.9
Economic & Soc Hist (620) - All	200	83.2	21.7	124	113	79	1012	2.7	7.9	62.3
Schools							82	4.9	11.0	69.5
FE							765	2.7	7.8	62.0
Overseas							0	0.0	0.0	0.0
Male							390	3.1	7.2	60.5
Female							622	2.4	8.4	63.3
Economics (618) - All	200	90.8	29.6	133	115	72	7331	8.8	22.9	72.6
Schools							3320	10.1	25.2	74.9
FE							3206	7.9	22.0	72.1
Overseas							69	2.9	7.2	49.3
Male							4817	8.9	23.3	74.2
Female							2514	8.6	22.1	69.6
English I (Lang & Lit) (623) - All	300	163.0	33.1	208	185	130	14774	9.0	25.7	85.3
Schools							2247	11.0	29.6	90.8
FE							10703	8.6	25.2	84.6
Overseas							38	13.2	21.1	73.7
Male							4281	9.6	24.5	82.0
Female							10493	8.8	26.2	86.6
English II (Literature) (652) - All	300	117.8	47.8	183	154	77	2367	10.1	22.8	80.3
Schools							525	12.4	24.8	84.2
FE							1374	9.1	21.9	77.8
Overseas							196	10.2	27.0	83.7
Male							678	10.6	21.4	76.4
Female							1689	9.8	23.4	81.8

1994

English III (Lit. Alt.) (660) - All	360	218.2	44.7	270	242	151	19698	13.9	30.7	94.7
Schools							10416	14.4	30.9	94.2
FE							7757	12.9	30.5	95.3
Overseas							20	20.0	65.0	100.0
Male							5598	14.2	30.8	93.8
Female							14100	13.8	30.7	95.0
Environ Science (657) - All	250	121.0	27.9	164	142	95	1699	7.2	24.5	84.5
Schools							590	8.0	27.1	85.1
FE							915	5.6	20.1	83.2
Overseas							48	16.7	41.7	93.8
Male							855	7.0	27.1	86.5
Female							844	7.5	21.9	82.5
French (624) - All	500	258.3	63.2	323	287	204	5345	16.7	33.6	80.1
Schools							2401	16.3	33.6	81.1
FE							2556	16.5	33.2	79.6
Overseas							16	25.0	68.8	93.8
Male							1442	18.4	36.6	83.3
Female							3903	16.1	32.5	78.9
General Studies (6667) - All	400	205.6	42.0	259	231	157	2305	10.6	27.2	88.9
Schools							1382	10.3	27.4	88.1
FE							585	11.3	26.8	91.1
Overseas							17	64.7	88.2	100.0
Male							1045	10.2	24.8	87.0
Female							1260	10.9	29.2	90.5
Geography (626) - All	600	271.1	60.7	337	299	225	3745	14.3	32.8	77.7
Schools							1967	15.0	34.5	80.2
FE							1421	14.8	32.0	74.5
Overseas							59	13.6	39.0	81.4
Male							1997	11.6	30.5	78.4
Female							1748	17.4	35.4	76.8

1994

German (629) - All	500	238.0	73.3	307	259	168	2187	18.0	37.7	82.9
Schools							1127	20.4	42.5	84.9
FE							956	15.4	32.4	80.8
Overseas							11	18.2	54.5	100.0
Male							664	20.6	41.7	84.6
Female							1523	16.9	35.9	82.1
Government Politics (0609B) All	320	182.0	47.7	239	209	139	918	12.0	32.2	82.1
Schools							436	13.3	34.6	83.7
FE							405	12.1	30.9	80.0
Overseas							1	0.0	0.0	100.0
Male							595	10.8	32.9	82.4
Female							323	14.2	31.0	81.7
History (0630BM) All	200	96.0	25.3	132	116	77	2515	7.4	23.3	78.8
Schools							948	11.7	30.9	83.9
FE							1437	5.1	19.5	76.3
Overseas							3	0.0	0.0	100.0
Male							1223	9.0	25.9	80.5
Female							1292	6.0	20.8	77.2
History (Alternative) (0673B) All	400	193.5	47.1	255	227	151	1638	9.7	24.8	82.7
Schools							1077	10.1	24.9	82.0
FE							415	9.6	26.3	83.9
Overseas							0	0.0	0.0	0.0
Male							681	10.7	27.5	83.4
Female							957	9.0	22.9	82.2
History of Art (606F) All	200	73.1	22.7	122	100	58	811	2.2	12.9	76.0
Schools							168	0.6	11.3	78.0
FE							566	2.5	13.4	73.9
Overseas							3	0.0	0.0	100.0
Male							210	2.9	11.4	74.3
Female							601	2.0	13.5	76.5

1994

Human Biology (642) - All	390	172.3	50.9	246	218	153	3956	9.0	19.6	63.3
Schools							344	11.0	24.7	68.0
FE							3350	8.8	19.2	63.1
Overseas							21	4.8	28.6	85.7
Male							1101	11.8	22.3	62.9
Female							2855	8.0	18.6	63.4
Law (625) - All	200	79.3	25.1	110	99	70	4592	10.8	23.0	66.7
Schools							398	11.1	25.6	69.6
FE							3554	10.9	23.2	66.2
Overseas							7	0.0	0.0	71.4
Male							1858	10.4	21.9	63.1
Female							2734	11.0	23.7	69.1
Music (633) - All	800	428.2	95.1	501	457	328	201	19.9	36.8	84.6
Schools							110	20.0	40.0	86.4
FE							66	19.7	28.8	81.8
Overseas							11	36.4	63.6	90.9
Male							72	12.5	31.9	88.9
Female							129	24.0	39.5	82.2
Philosophy (661) - All	200	93.4	32.9	137	116	72	967	9.6	28.0	74.7
Schools							201	15.9	39.3	86.6
FE							679	6.8	23.3	71.3
Overseas							2	0.0	50.0	100.0
Male							503	9.3	27.2	73.6
Female							464	9.9	28.9	75.9
Photography (634) - All	300	152.2	29.7	189	166	121	1578	11.7	32.3	85.7
Schools							166	3.6	20.5	77.1
FE							1221	12.9	33.5	86.1
Overseas							1	0.0	100.0	100.0
Male							892	13.1	33.7	87.1
Female							686	9.9	30.5	84.0

1994

Physics A (635) - All	500	283.7	81.8	372	331	228	3589	16.6	31.7	76.0
Schools							1468	21.2	37.1	68.9
FE							1474	12.7	26.5	80.5
Overseas							302	16.2	34.1	71.8
Male							2806	16.7	31.5	75.1
Female							783	16.1	32.1	72.5
Physics B (635) - All	500	275.7	80.5	365	327	227	1709	15.2	28.4	71.7
Schools							512	18.0	30.7	71.9
FE							778	12.3	24.8	69.7
Overseas							247	17.0	34.0	78.9
Male							1353	15.5	28.6	71.5
Female							356	14.0	27.5	72.5
Psychology (651) - All	500	244.4	68.6	331	292	193	18537	10.3	26.2	77.2
Schools							2470	10.7	25.8	78.8
FE							14514	10.3	26.3	77.0
Overseas							140	14.3	31.4	79.3
Male							4484	7.6	20.8	70.6
Female							14053	11.1	27.9	79.2
Pure & Applied Maths (636) - All	210	103.3	46.9	159	135	80	4049	14.1	27.8	68.1
Schools							1872	18.5	33.8	76.5
FE							1772	10.4	23.7	60.9
Overseas							46	10.9	28.3	71.7
Male							2988	13.0	26.1	66.1
Female							1061	17.0	32.5	73.7
Pure Mathematics (632) - All	210	119.5	52.5	161	144	86	1682	26.3	38.9	73.1
Schools							487	45.8	61.0	87.5
FE							544	23.2	38.4	74.3
Overseas							464	11.0	20.0	62.1
Male							1130	27.2	39.5	73.8
Female							552	24.5	37.7	71.6

1994

Pure Maths & Stats (646) - All	210	107.0	42.2	162	141	86	3192	9.9	23.9	68.9
Schools							1254	12.3	28.3	76.4
FE							1666	8.5	21.7	64.9
Overseas							32	6.3	12.5	43.8
Male							1559	9.6	23.5	65.8
Female							1633	10.1	24.4	71.9
Sociology (639) - All	200	85.9	24.1	119	106	79	17036	8.3	21.2	63.6
Schools							4845	10.7	25.6	71.2
FE							10361	7.5	19.7	61.2
Overseas							158	5.7	13.9	49.4
Male							4592	7.4	18.8	59.1
Female							12444	8.6	22.0	65.3
Sociology (664) - All	200	100.0	21.8	128	112	87	9372	10.3	30.0	73.6
Schools							3364	10.3	30.7	73.5
FE							5392	10.1	29.5	73.4
Overseas							35	14.3	22.9	77.1
Male							2260	8.2	26.2	68.7
Female							7112	10.9	31.3	75.2
Spanish (640) - All	500	261.0	68.4	317	283	193	1054	20.7	39.7	83.4
Schools							343	22.2	44.3	86.3
FE							613	18.1	35.4	81.1
Overseas							5	100.0	100.0	100.0
Male							304	27.0	45.4	89.1
Female							750	18.1	37.3	81.1
Sport Studies (665) - All	250	106.6	21.9	147	129	82	2015	4.1	17.2	87.8
Schools							505	6.5	20.0	90.9
FE							1307	3.5	16.8	87.1
Overseas							0	0.0	0.0	0.0
Male							1281	2.9	16.2	87.7
Female							734	6.3	18.9	87.9

1994

Statistics (641) - All	210	120.0	45.5	165	146	87	1062	16.5	34.6	76.3
Schools							217	27.6	46.1	86.6
FE							515	9.3	25.8	72.0
Overseas							245	21.2	43.3	79.6
Male							638	17.9	37.5	78.8
Female							424	14.4	30.2	72.4
Theatre Studies (653) - All	250	129.0	28.2	166	150	105	7789	10.3	24.1	80.6
Schools							3912	11.7	26.3	82.4
FE							3419	8.7	21.6	78.8
Overseas							11	11.0	18.2	63.6
Male							2060	8.1	19.1	74.0
Female							5729	11.1	25.9	83.0
Physical Education (703) - All	1000	529.4	93.7	658	601	405	3621	9.0	23.2	91.5
Schools							1322	10.5	24.7	92.8
FE							2160	7.9	22.2	90.7
Overseas							0	0.0	0.0	0.0
Male							2357	6.9	20.5	91.3
Female							1264	12.9	28.2	92.0

APPENDIX 6.6

THE EFFECTS OF A CHANGING COMPOSITION OF CANDIDATES UPON THE CUMULATIVE PROPORTION OF CANDIDATES AT ANY GIVEN GRADE

The candidates as a whole are divided into n subgroups such as different centre types and/or genders.

Let the proportion of the candidates in Year 1 belonging to subgroup i be p_i

the overall cumulative proportion of candidates at an arbitrary grade boundary in Year 1 be C_1

the difference between C_1 and the cumulative proportion of candidates at the arbitrary grade boundary for subgroup i be δ_i

It follows, summing across the subgroups, that:

$$\sum_{i}^{n} (p_i) = 1 \qquad \text{and} \qquad \sum_{i}^{n} (p_i d_i) = 0 \qquad \text{[Equations 1 \& 2]}$$

Now let the change in the proportion of the candidates belonging to subgroup i between Year 1 and Year 2 be δ_i

Clearly, $$\sum_{i}^{n} (\delta_i) = 0 \qquad \text{[Equation 3]}$$

If it is now assumed that the subgroups remain the same in terms of their relative achievement (that is that the d_i remain the same from Year 1 to Year 2) then the cumulative proportion of candidates at the arbitrary grade boundary in Year 2 is given by:

$$C_2 = \sum_{i}^{n} (p_i + \delta_i)(C_1 + d_i)$$

Equations 1, 2 and 3 reduce this to:

$$C_2 = C_1 + \sum_{i}^{n} (\delta_i d_i)$$

Thus, the change in the overall cumulative proportion at the arbitrary boundary between Years 1 and 2 is the sum, across the subgroups, of the products of two terms:

the change in incidence of each subgroup; and

the difference between the cumulative proportion for the subgroup and the overall cumulative proportion.

In practice, both of these terms are normally small (<< 0.5).

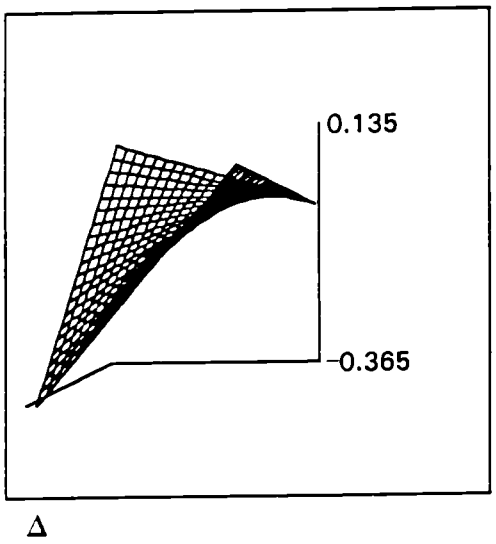
In the case of two subgroups, the term for the change in the cumulative proportion can be modelled as the height of a surface in a three dimensional space in which the horizontal plane is defined by δ_1 and d_1. Four examples which use this approach are shown below. The response surface is saddle-shaped, as expected it is fairly flat and has little height for wide ranges of plausible values of δ_1 and d_1.

EXAMPLE 1

$p_1 = 0.2$ $C_1 = 0.73$

$$\Delta_{i,j} = \delta 1_i \cdot d1_j + \delta 2_i \cdot d2_j$$

$\delta 1_i$	$\delta 2_i$
-0.1	0.1
-0.075	0.075
-0.05	0.05
-0.025	0.025
0	0
0.025	-0.025
0.05	-0.05
0.075	-0.075
0.1	-0.1
0.125	-0.125
0.15	-0.15
0.175	-0.175
0.2	-0.2
0.225	-0.225
0.25	-0.25
0.275	-0.275
0.3	-0.3
0.325	-0.325
0.35	-0.35
0.375	-0.375
0.4	-0.4



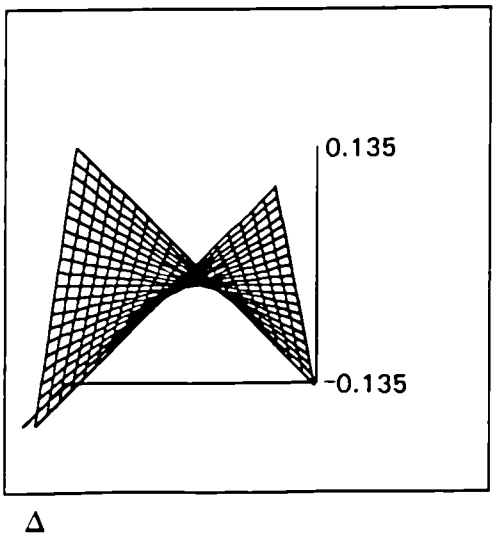
$d1_j$	$d2_j$
-0.73	0.183
-0.68	0.17
-0.63	0.158
-0.58	0.145
-0.53	0.133
-0.48	0.12
-0.43	0.108
-0.38	0.095
-0.33	0.082
-0.28	0.07
-0.23	0.058
-0.18	0.045
-0.13	0.033
-0.08	0.02
-0.03	0.008
0.02	-0.005
0.07	-0.018
0.12	-0.03
0.17	-0.043
0.22	-0.055
0.27	-0.068

EXAMPLE 2

$p_1 = 0.5$ $C_1 = 0.73$

$$\Delta_{i,j} = \delta 1_i \cdot d1_j + \delta 2_i \cdot d2_j$$

$\delta 1_i$	$\delta 2_i$
-0.25	0.25
-0.225	0.225
-0.2	0.2
-0.175	0.175
-0.15	0.15
-0.125	0.125
-0.1	0.1
-0.075	0.075
-0.05	0.05
-0.025	0.025
0	0
0.025	-0.025
0.05	-0.05
0.075	-0.075
0.1	-0.1
0.125	-0.125
0.15	-0.15
0.175	-0.175
0.2	-0.2
0.225	-0.225
0.25	-0.25



$d1_j$	$d2_j$
-0.27	0.27
-0.243	0.243
-0.216	0.216
-0.189	0.189
-0.162	0.162
-0.135	0.135
-0.108	0.108
-0.081	0.081
-0.054	0.054
-0.027	0.027
0	0
0.027	-0.027
0.054	-0.054
0.081	-0.081
0.108	-0.108
0.135	-0.135
0.162	-0.162
0.189	-0.189
0.216	-0.216
0.243	-0.243
0.27	-0.27

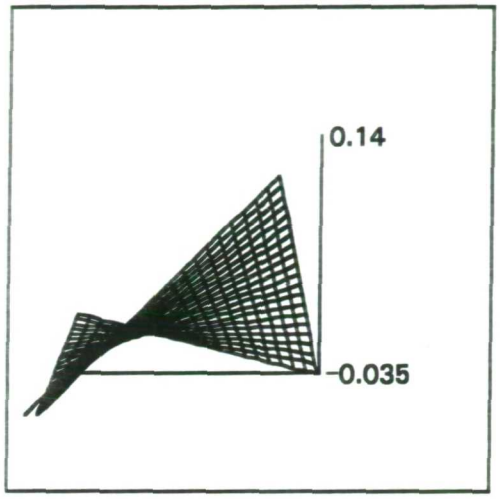
EXAMPLE 3

$p_1 := 0.2$ $C_1 := 0.07$

$\Delta_{i,j} := \delta 1_i \cdot d1_j + \delta 2_i \cdot d2_j$

$\delta 1_i$	$\delta 2_i$
-0.1	0.1
-0.075	0.075
-0.05	0.05
-0.025	0.025
0	0
0.025	-0.025
0.05	-0.05
0.075	-0.075
0.1	-0.1
0.125	-0.125
0.15	-0.15
0.175	-0.175
0.2	-0.2
0.225	-0.225
0.25	-0.25
0.275	-0.275
0.3	-0.3
0.325	-0.325
0.35	-0.35
0.375	-0.375
0.4	-0.4

$d1_j$	$d2_j$
-0.07	0.018
-0.053	0.013
-0.035	0.009
-0.018	0.004
0	0
0.017	-0.004
0.035	-0.009
0.052	-0.013
0.07	-0.017
0.087	-0.022
0.105	-0.026
0.123	-0.031
0.14	-0.035
0.158	-0.039
0.175	-0.044
0.192	-0.048
0.21	-0.052
0.228	-0.057
0.245	-0.061
0.263	-0.066
0.28	-0.07



Δ

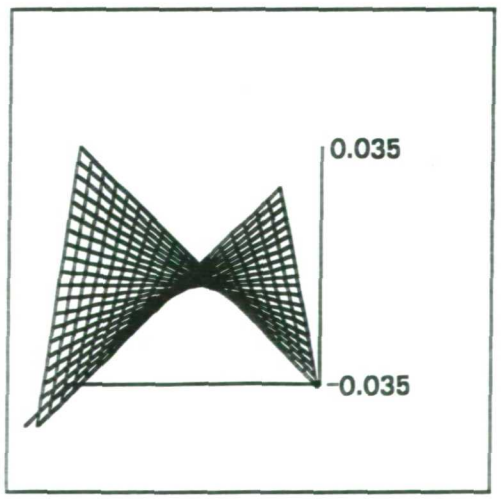
EXAMPLE 4

$p_1 := 0.5$ $C_1 := 0.07$

$\Delta_{i,j} := \delta 1_i \cdot d1_j + \delta 2_i \cdot d2_j$

$\delta 1_i$	$\delta 2_i$
-0.25	0.25
-0.225	0.225
-0.2	0.2
-0.175	0.175
-0.15	0.15
-0.125	0.125
-0.1	0.1
-0.075	0.075
-0.05	0.05
-0.025	0.025
0	0
0.025	-0.025
0.05	-0.05
0.075	-0.075
0.1	-0.1
0.125	-0.125
0.15	-0.15
0.175	-0.175
0.2	-0.2
0.225	-0.225
0.25	-0.25

$d1_j$	$d2_j$
-0.07	0.07
-0.063	0.063
-0.056	0.056
-0.049	0.049
-0.042	0.042
-0.035	0.035
-0.028	0.028
-0.021	0.021
-0.014	0.014
-0.007	0.007
0	0
0.007	-0.007
0.014	-0.014
0.021	-0.021
0.028	-0.028
0.035	-0.035
0.042	-0.042
0.049	-0.049
0.056	-0.056
0.063	-0.063
0.07	-0.07



Δ

ACKNOWLEDGEMENT

I am grateful to Simon Eason for the original algebraic formulation upon which this appendix is based.

APPENDIX 6.7

FORM USED BY OBSERVERS TO ENCODE AWARDERS' CONTRIBUTIONS DURING PHASE 3

Type	Chair	Awarders	Officers
Affective/ Social	Positive	Positive	Positive
	Negative	Negative	Negative
Methodological/ Procedural suggestion	Gives	Gives	Gives
	Seeks	Seeks	Seeks
Criteria for evaluating scripts	Gives	Gives	Gives
	Seeks	Seeks	Seeks
Overall judgement of candidates	Gives	Gives	Gives
	Seeks	Seeks	Seeks
Evaluation of script(s)	Gives	Gives	Gives
	Seeks	Seeks	Seeks
Statistical opinion/information	Gives	Gives	Gives
	Seeks	Seeks	Seeks
Paper or Question opinion/information	Gives	Gives	Gives
	Seeks	Seeks	Seeks
Boundary suggestion	Gives	Gives	Gives
	Seeks	Seeks	Seeks
Other			

APPENDIX 6.8

QUANTITATIVE DATA FROM PHASE 3 OBSERVATIONS

Direction	Nature of remark	Economics Total All Papers and Boundaries						Total
		Observer 1			Observer 2			
		Chair	Awarders	Officers	Chair	Awarders	Officers	
Seeks	Other relevant	0	0	0	0	0	0	**0**
	Paper/task opin/info	5	1	1	4	0	1	**12**
	Statistical opin/info	4	2	0	5	2	0	**13**
	Script evaluation	6	1	0	6	3	0	**16**
	Overall judgement	0	0	0	0	0	0	**0**
	Evaluative criterion	1	1	1	1	1	1	**6**
	Methodological/Procedural	6	2	1	6	2	1	**18**
	Affective	0	0	0	0	0	0	**0**
	Boundary suggestion	7	7	0	6	5	0	**25**
Gives	Boundary suggestion	15	9	6	16	13	6	**65**
	Affective	20	5	0	22	4	0	**51**
	Methodological/Procedural	9	6	3	9	7	4	**38**
	Evaluative criterion	5	11	1	6	9	1	**33**
	Overall judgement	2	2	1	2	2	1	**10**
	Script evaluation	26	39	1	30	42	0	**138**
	Statistical opin/info	8	5	20	7	4	23	**67**
	Paper/task opin/info	13	20	1	9	13	1	**57**
	Other relevant	0	0	0	0	0	0	**0**
Total		**127**	**111**	**36**	**129**	**107**	**39**	**549**

Direction	Nature of remark	English Total All Papers and Boundaries						Total
		Observer 1			Observer 2			
		Chair	Awarders	Officers	Chair	Awarders	Officers	
Seeks	Other relevant	0	0	0	0	0	0	**0**
	Paper/task opin/info	1	8	0	1	8	0	**18**
	Statistical opin/info	11	10	0	9	16	1	**47**
	Script evaluation	15	11	1	17	21	0	**65**
	Overall judgement	0	0	0	0	0	0	**0**
	Evaluative criterion	0	0	0	0	2	0	**2**
	Methodological/Procedural	2	1	0	1	5	0	**9**
	Affective	0	1	0	0	1	0	**2**
	Boundary suggestion	20	9	5	23	12	2	**71**
Gives	Boundary suggestion	31	83	23	34	79	21	**271**
	Affective	31	24	2	37	27	3	**124**
	Methodological/Procedural	5	2	1	8	8	2	**26**
	Evaluative criterion	2	3	0	3	9	0	**17**
	Overall judgement	3	5	0	5	8	0	**21**
	Script evaluation	71	148	5	72	153	0	**449**
	Statistical opin/info	14	31	49	17	38	50	**199**
	Paper/task opin/info	14	28	3	6	21	5	**77**
	Other relevant	0	0	0	0	0	0	**0**
Total		**220**	**364**	**89**	**233**	**408**	**84**	**1398**

Direction	Nature of remark	Mathematics Total All Papers and Boundaries						Total
		Observer 1			Observer 2			
		Chair	Awarders	Officers	Chair	Awarders	Officers	
Seeks	Other relevant	0	0	0	0	0	0	**0**
	Paper/task opin/info	3	21	1	0	29	2	**56**
	Statistical opin/info	26	24	0	28	36	0	**114**
	Script evaluation	4	1	0	7	2	0	**14**
	Overall judgement	0	0	0	1	0	0	**1**
	Evaluative criterion	0	0	0	0	2	0	**2**
	Methodological/Procedural	6	5	3	6	7	8	**35**
	Affective	0	0	0	0	0	0	**0**
	Boundary suggestion	31	9	6	33	11	7	**97**
Gives	Boundary suggestion	33	67	23	33	77	20	**253**
	Affective	11	33	0	10	38	1	**93**
	Methodological/Procedural	11	12	16	14	25	18	**96**
	Evaluative criterion	3	15	0	4	26	3	**51**
	Overall judgement	3	26	2	3	30	2	**66**
	Script evaluation	33	138	2	28	154	2	**357**
	Statistical opin/info	51	84	54	45	92	65	**391**
	Paper/task opin/info	5	79	4	6	64	6	**164**
	Other relevant	0	0	0	0	0	0	**0**
Total		**220**	**514**	**111**	**218**	**593**	**134**	**1790**

Direction	Nature of remark	Physics Total All Papers and Boundaries						Total
		Observer 1			Observer 2			
		Chair	Awarders	Officers	Chair	Awarders	Officers	
Seeks	Other relevant	0	0	0	0	0	0	**0**
	Paper/task opin/info	0	1	0	1	1	0	**3**
	Statistical opin/info	18	3	0	12	4	0	**37**
	Script evaluation	3	1	0	3	0	0	**7**
	Overall judgement	0	0	0	0	0	0	**0**
	Evaluative criterion	0	0	0	1	1	0	**2**
	Methodological/Procedural	0	1	0	4	2	0	**7**
	Affective	2	0	0	0	0	0	**2**
	Boundary suggestion	12	5	1	10	1	0	**29**
Gives	Boundary suggestion	26	37	1	10	15	0	**89**
	Affective	63	27	0	37	22	1	**150**
	Methodological/Procedural	8	1	1	4	1	1	**16**
	Evaluative criterion	7	8	1	0	5	0	**21**
	Overall judgement	4	7	0	5	6	0	**22**
	Script evaluation	22	36	1	7	21	0	**87**
	Statistical opin/info	26	28	29	10	15	11	**119**
	Paper/task opin/info	7	31	0	4	8	1	**51**
	Other relevant	0	0	0	0	0	0	**0**
Total		**198**	**186**	**34**	**108**	**102**	**14**	**642**

Direction	Nature of remark	Communication studies Total All Papers and Boundaries						Total
		Observer 1			Observer 2			
		Chair	Awarders	Officers	Chair	Awarders	Officers	
Seeks	Other relevant	0	0	0	0	0	0	**0**
	Paper/task opin/info	1	0	0	0	4	0	**5**
	Statistical opin/info	2	0	0	3	2	0	**7**
	Script evaluation	8	1	0	5	0	0	**14**
	Overall judgement	0	0	0	0	0	0	**0**
	Evaluative criterion	0	0	0	0	0	0	**0**
	Methodological/Procedural	2	0	0	1	0	0	**3**
	Affective	0	0	0	0	0	0	**0**
	Boundary suggestion	2	0	0	3	1	0	**6**
Gives	Boundary suggestion	5	5	4	6	5	3	**28**
	Affective	2	0	0	7	8	0	**17**
	Methodological/Procedural	4	3	1	3	4	2	**17**
	Evaluative criterion	0	1	0	0	0	0	**1**
	Overall judgement	1	0	0	2	2	0	**5**
	Script evaluation	5	14	0	6	17	0	**42**
	Statistical opin/info	6	3	8	4	5	8	**34**
	Paper/task opin/info	0	2	0	1	8	1	**12**
	Other relevant	0	0	0	0	0	0	**0**
Total		**38**	**29**	**13**	**41**	**56**	**14**	**191**

APPENDIX 6.9

DATA ON AGREEMENT BETWEEN AUTHOR AND PHASE 3 OBSERVERS

		Economics P3 A/B boundary								
		Observer 1			Observer 2			Author		
		Chair	Awarders	Officers	Chair	Awarders	Officers	Chair	Awarders	Officers
Seeks	Other relevant									
	Paper/task opin/info	3			2			2		
	Statistical opin/info							1		
	Script evaluation	1			2	1		3	3	
	Overall judgement									
	Evaluative criterion		1			1				
	Methodological/Procedural	1	2		1	1				
	Affective									
	Boundary suggestion	2	1		2	1		4	2	
Gives	Boundary suggestion	8	2	1	8	3	1	6	2	2
	Affective	7	4		7	3		3		
	Methodological/Procedural	2	1		2	1		1		
	Evaluative criterion	3	5		2	3			4	
	Overall judgement		1			1				
	Script evaluation	7	13	1	9	14		6	15	
	Statistical opin/info	2	1	3	2	1	4	2	1	3
	Paper/task opin/info	4	11		4	9		3	9	
	Other relevant									

		Economics P3 B/C boundary								
		Observer 1			Observer 2			Author		
		Chair	Awarders	Officers	Chair	Awarders	Officers	Chair	Awarders	Officers
Seeks	Other relevant									
	Paper/task opin/info									
	Statistical opin/info	2			2			1		
	Script evaluation	2	1		1	1		2	1	
	Overall judgement									
	Evaluative criterion									
	Methodological/Procedural									
	Affective									
	Boundary suggestion	1	1		2			2	1	
Gives	Boundary suggestion	3	1	2	3	2	3	3	3	2
	Affective	2	1		3	1		3		
	Methodological/Procedural	2	1	1	1	1	2			2
	Evaluative criterion				1	1		1		
	Overall judgement	1		1	1		1			1
	Script evaluation	2	5		3	6	0	2	7	
	Statistical opin/info	1	1	4	1	1	4	1	1	4
	Paper/task opin/info	1						1		
	Other relevant									

		Economics P3 E/N boundary								
		Observer 1			Observer 2			Author		
		Chair	Awarders	Officers	Chair	Awarders	Officers	Chair	Awarders	Officers
Seeks	Other relevant									
	Paper/task opin/info		1					2		
	Statistical opin/info	1	1		2	1		2	1	
	Script evaluation	3			3			3		
	Overall judgement									
	Evaluative criterion									
	Methodological/Procedural	2			2	1		2		
	Affective									
	Boundary suggestion	2	2		2	1		2	2	
Gives	Boundary suggestion	1	3	1	2	3	1	3	3	1
	Affective	4			4			3		
	Methodological/Procedural	2	1	1	2	2	1		2	1
	Evaluative criterion		2			2		1	3	
	Overall judgement		1			1				
	Script evaluation	4	10		5	8		4	10	
	Statistical opin/info	2	1	4	2	1	4	2	2	4
	Paper/task opin/info	2	2			1			2	
	Other relevant									

APPENDIX 7.1

WESSEX PROJECT AWARDING PROCEDURES

THE AWARD OF CORE LEVELS, AND THE VERIFICATION OF GRADES FOR THE WESSEX PROJECT

1. INTRODUCTION

1.1 This paper is concerned with two processes which are designed to occur in rapid succession at the end of the two year period of the normal Wessex A Level course; these are the core award and the verification procedure. Most of the procedures described here are presented in a way which makes them applicable over a wider range of subjects, although they will need to be supplemented in order to deal with components in which assessments have been made directly into levels, and where some special types of work need to be scrutinised.

1.2 For convenience the paper deals separately with the core award and verification procedures. Although not essential, it is desirable that these take place on the same day, and with the same team of people. To achieve this will require a degree of forward planning, but the procedures are unlikely to succeed unless they can be fitted alongside existing award commitments.

1.3 This paper is written for the information and guidance of the subject officer and other Board staff. Attached to it are Appendices which can be supplied to participants in the core award and verification meetings, as self-contained procedure documents. Additional materials, as listed in the following sections, will also be made available.

2. ARRANGING THE MEETINGS

2.1 It is assumed that some of the material in the core examination will be identical, or very similar, to some of that used in the current Mode 1 examination in the subject. For this reason it is intended that the award of core levels shall normally be done on the same occasion as the award of the appropriate Mode 1 examination, and always involve some of the same awarders. The extent to which it is possible to extend the Mode 1 award meeting in order to include the Wessex core award on the same day is a decision to be made by the subject officer, in the knowledge of the workload involved in the Mode 1. It is probable that the awarders will need to continue on a second day, and will then follow the Wessex core award with the verification meeting. In all cases where the Wessex core draws on the Mode 1 examination for material it should follow and not precede the Mode 1 award. The verification meeting then comes last.

2.2 Although it is not essential that all of the Wessex core awarders have been involved in the relevant Mode 1 award it is very desirable that most of them have been. The purpose of a common membership is to ensure some carry-over of standards from one meeting to the other, and this can normally be achieved by having some of the Mode 1 awarders staying on for the Wessex core award, and then for the verification meeting. The Wessex core award meeting should have the following membership:

Two or three members of the AEB Standing Advisory Committee (SAC) of whom at least two have been at the Mode 1 award;

The subject co-ordinator from the Wessex Project, or an alternate who has the required subject expertise;

The chief examiner.

The same personnel will conduct the verfication procedure, with the addition of the moderator who has dealt with the modules. (There is no reason why the moderator

should not observe the core award if it is convenient.) The subject officer will service both meetings, and other Board staff may attend in order to observe or advise as necessary.

2.3 A chairman will be appointed for each meeting from amongst the SAC members; it will often be convenient for this to be the same person for both, and often the same person as chaired the Mode 1 award. Because of the need to allow sufficient time for the verification process, the chairman must see that the award meeting proceeds reasonably quickly: this should be most easily possible when the overlap with the Mode 1 examination is substantial.

3. PRELIMINARY INFORMATION AND MATERIALS AT THE MEETINGS

3.1 The structure of the Wessex scheme is not the same as that of Mode 1 examinations, and the syllabus of a Wessex subject is not the same as that for a Mode 1. It is therefore essential that those who are attending the Wessex meetings are provided with adequate background information, presented in a way which makes clear what they are dealing with. [1] The subject officer will therefore supply to the awarders, in advance of the meeting day, a general package of information, with one supplementary package relating to the core award, and a second supplementary package relating to the verification procedure.

3.2 The general package will include

- a brief introduction to the general features of the Wessex scheme, in terms which relate directly to the subject concerned (see Appendix 1)
- a syllabus for the subject concerned

[1] In future it will be possible to treat a core award for this scheme very like a conventional award, supplying past scripts, etc. as described in AEB Procedure File 7: Guidelines for the Conduct of AEB Advanced Level Grade Awarding Meetings. In the meantime a rather reduced amount of material can be made available.

- the normal claim forms etc.

3.3 The core award package will include

- a paper describing the procedure to be followed, drawing particular attention to the place of the core award within the structure of the scheme (see Appendix 2)
- copies of core written papers and mark schemes
- copies of materials relating to non-written core components (if these exist)
- (after the first year) materials relating to previous core awards.

3.4 The verification package will include

- a description of the procedure to be followed (see Appendix 3)
- (where the personnel concerned have not already had the materials) copies of Wessex written papers and other assessment materials and copies of Board examination question papers etc.

3.5 For the core award the subject officer will arrange for the same materials to be on hand as for a Mode 1 award, namely

- all scripts, arranged by centre within core component (see Appendix 2)
- mark distributions for core components and for the core as a whole (see Appendix 2)
- form for recording awarding decisions (see Appendix 4)

3.6 The subject officer will arrange for the collection of materials from the modules, so that they can be scrutinised at the verification meeting. The amount of material to be made available may vary somewhat amongst the subjects concerned, and may be partly conditioned by the nature of the work done for the modules. Whilst the scheme is in its

early pilot stages Board officers will arrange for all, or a large proportion of, module reports to be brought to the Board, and to these will be added distributions of marks and levels awarded on each of the modules.

3.7 The verification meeting can only proceed if the whole work of candidates can be scrutinised, using the procedures described in Appendix 3. In order for this to be accomplished, and to avoid wasting time in assembling materials during the meeting, the module material from the candidates concerned will be put together, by candidate, in advance of the meeting, and a form attached which shows

- the centre number
- the candidate number
- the level achieved on each module

and upon which there are spaces for the core results, and the A Level grade to be inserted (see Appendix 4). Following the core award, the subject officer will arrange for a clerk to add the core script(s) and record of any other core results (such as that from a practical component) to the pile of each candidate's work, for the core result to be inserted on the form, and for the A Level grade (calculated using the combination rules) to be written down. The participants in the verification meeting will thus be able to have immediate access to all of the information required in order for them to follow the procedure described in Appendix 3.

3.8 In addition to the script and module material available at the verification meeting forms which enable the participants to record their opinions, and to describe the outcomes of their discussions are also needed; these are shown in Appendix 4. It is important that the subject officer collects and files these forms with those from the core award, since all of this information will be needed for the evaluation of the Wessex scheme.

3.9 The responsibilities of the participants in both meetings include the need for complete familiarity with all of the materials which have been sent out in advance. Additionally, the

chief examiner and moderator will bring to the meetings particular information and expertise as specified in the meeting procedure papers in Appendices 2 and 3. It is essential that all of those concerned fully appreciate the differences between the two meetings, and how both relate to the Mode 1 award with which some of them are also concerned. It is inevitable that these two tasks will be especially onerous on this first occasion, and the participants will need to be patient, and to be prepared to contribute to the development of the Project.

3.10 The outcomes of both of these meetings remain confidential until results are published by the Board. In due course arrangements may be possible whereby core results can be issued separately, but this might involve a departure from the inter-Board agreement on a single results issue date. Until such time as this is clarified, and the Wessex procedures are approved (that is, have been piloted, evaluated and agreed to), the results which come from the core award and verification procedures will not be made available before all other A Level results.

J. Wilmut/W.A.R. Gardiner

May 1990

WESSEXCO

APPENDIX 1

GENERAL INTRODUCTION TO THE WESSEX SCHEME

The Wessex Project is a collaborative venture involving the education authorities of Avon, Dorset, Gloucestershire, Somerset and Wiltshire and the Associated Examining Board. The Project is concerned with improving education provision post-sixteen through the introduction of modular courses at A and AS Level and the creation of a Module "Bank" which will have even wider applications.

All Modular A Level courses consist of a Core (60%) and four Modules (4 x 10%). AS Levels will consist of Core (60%) and two Modules (2 x 20%). Cores, which allow for continuity over two years, provide for a solid foundation of knowledge, skills and understanding. All Cores include external assessment. Modules provide opportunities for cross-curricular activities, student-centred work, greater flexibility and choice. Many Modules are inter-disciplinary in nature and focus on topical issues and themes while others are subject-specific. They are assessed as coursework.

It is through a combination of active learning and student-centred approaches, relevant courses of study, academic tutoring, records of achievement, and a modular structure that Wessex aims to bridge the academic-vocational divide without loss of rigour or coherence in Advanced Level studies.

The Project attempts to

- provide continuity with GCSE through its assessment and teaching styles
- broaden the base of A Levels by emphasising skills and experiences
- create modules which cross traditional subject boundaries
- provide flexibility of accreditation
- incorporate student profiling
- respond to student needs.

The Project started in 1986 and students began the A Level Chemistry course in 1987. Physics, Biology, Design and Technology, French and German followed in 1988. All of these have been approved for pilot status by the School Examinations and Assessment Council (SEAC) and have limits on the number of students who may start them.

The same pilot status has now been given to A Level syllabuses in Geography, Business Studies & Economics, and Art & Design - which started in 1989. Development work has also begun in English, Mathematics, Performing Arts, Home Economics and History and there are plans for the introduction of AS Level syllabuses as soon as possible.

All development has been undertaken by teachers in the Wessex authorities, advisers and AEB personnel. Financial support has been provided by the LEAs, supplemented by a considerable amount of commercial and industrial sponsorship. Each subject development team has its own co-ordinator responsible for constructing the syllabus, co-ordinating course delivery and producing learning materials.

The Project is being evaluated in two ways. The Training Agency of the Department of Employment is providing the funding for a three year evaluation of the scheme under the direction provided by an independent evaluator from the School of Education at Bath University. AEB is also responsible for an evaluation of the assessment procedures and examination results, in order to report to SEAC.

The views of Higher Education concerning Wessex syllabuses have been sought. The responses so far are encouraging. Towards the end of the pilot, revised syllabuses will be produced in readiness for the general availability of the scheme.

Modules share similar assessment objectives. The relationship between these and the Core objectives, with details of the Modules currently available, are given in the syllabus and Handbook.

There are three types of Modules:

A - subject-specific

B - subject-specific, but with overlap with another subject

C - interdisciplinary.

Each syllabus requires that students must select

at least 2 type A Modules

not more than 2 type B Modules

not more than 1 type C Module.

Modules may be taken at any time after term 1 which is devoted to delivery of the Core, although the teachers in the pilot institutions have pursued a policy of co-ordinated module delivery. This is assisted by the organisation of centres into local area groups called Networks. A Network provides the vehicle for INSET delivery and moderation. Each Subject Moderator is attached to a Network and moderates all subject-specific modules designated to that subject. (All Inter-disciplinary modules are designated subject-specific for one syllabus.)

A single procedure is being used for the generation of A Level grades for all subjects. It is being evaluated as part of the pilot, and the results will be reported to SEAC. Its first operational was at the Chemistry award in the summer of 1989.

Both Modules and Core have their results expressed in levels, which are reported to students. There are 6 levels for reporting the results on the Core, with an unclassified (U) category; level 1 is the highest. These levels will be awarded judgementally by an awarding panel. Module results are reported in 4 levels, with an unclassified (U) category; again level 1 is the highest. These levels are related directly to the marks whose award is controlled by the criteria shown in Appendix III of the syllabus.

Levels are recorded by the Board. Once a candidate has completed a permitted combination of Core and Modules, an A Level result will be generated using the following table. This table relates the grade obtained to the Core level gained (on the left) and to the sum of the Module levels (along the top). An unclassified result on the Core prohibits the award of an A Level grade. An unclassified Module does not prohibit such an award, so the table includes some A level awards which include one or two unclassified Module results.

Module Sum

Core Level	4	5	6	7	8	9	10	11 (3+U)	12	13 (4+U)	14	15 (5+U)	16 (2+2U)	6+U (3+2U)	7+U (4+2U)	8+U (5+2U)	9+U	10+U	11+U	12+U
1	A	A	A	A	A	A	B	B	B	B	C	C	C	C	D	D	E	E	N	N
2	A	A	B	B	B	B	B	C	C	C	C	D	D	D	D	E	E	N	N	U
3	B	B	B	C	C	C	C	C	D	D	D	D	D	E	E	N	N	U	U	U
4	C	C	C	C	D	D	D	D	D	E	E	E	E	N	N	U	U	U	U	U
5	D	D	D	D	D	E	E	E	E	E	N	N	N	U	U	U	U	U	U	U
6	D	D	E	E	E	E	E	N	N	N	N	N	N	U	U	U	U	U	U	U
U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U

It is because of the provisional nature of this procedure that a verification meeting will be held in order to determine whether the A Level grade outcomes are appropriate to the work that has been done by the candidates. This meeting will have available all the work done by at least a sample of the candidates, and will proceed, on a judgemental basis, to determine grades. These can be compared with those given by the procedure above; deficiencies in the procedure can then be remedied.

May 1990

[redacted]

APPENDIX 2

THE CONDUCT OF THE CORE AWARD

1. INTRODUCTION

1.1 This document should be read in conjunction with the General Introduction to the Wessex Scheme which describes the purpose and structure of the Scheme, and the piloting which is currently being undertaken. Those appointed to act as awarders for a Wessex core should also be in receipt of the appropriate syllabus, with which they should familiarise themselves, and may also receive other related materials designed to assist in the awarding process.

1.2 Unlike conventional A Level subjects there is no single grade awarding procedure for the Wessex Scheme. Instead there is a series of separate awards of levels, one award for each module, and a core award. It is the latter with which this document is concerned. Each Wessex A Level subject has a core which contributes 60% to the asssessment, leading to the A Level grade. The General Introduction indicates how this operates, providing a rationale for the core, and for the modules which are added to it. Additionally, the syllabus for the subject concerned will provide a detailed description of the core with which a particular core award meeting is concerned, describing the components into which it is divided, the objectives for each and so on.

1.3 The awarders are responsible for the determination of the positions of the boundaries between adjacent levels on the mark scale for the core. The procedures for doing this are described in detail below. They must accept that the marking of core components has been completed correctly and that each candidate's aggregate mark is the best available measure of his or her performance in the core as a whole. The process of determining the level boundaries is designed to be one where candidates' work is related to statements

which describe what is required for each level, and the cut-off points between levels are thereby determined. As the Wessex Project proceeds this process will refine through the re-interpretation of the descriptions of levels in the light of experience, and it is intended that the standard associated with each level for each core will remain stable from year to year.

1.4 However, it is emphasised that the Wessex Scheme is in its pilot phase, and that the first core award will be conducted without the benefit of prior experience, and with the need to establish the standards for the levels. Later sections of this document describe the procedures which will be used, but the innovative nature of the Scheme places an additional responsibility on awarders, who will be contributing to its development.

1.5 Generally speaking the people who conduct the core award will also participate in the verification process, which is the subject of a separate procedural document. It will be seen there that the outcome of the core and module awards is capable of modification following an overview. This procedure is, therefore, also contributing to the establishment of a stable interpretation of the core level standards, so that stability should emerge over a few awarding sessions.

1.6 The people present at the core award (SAC members, chief examiner(s), Wessex representative and subject officer) all have expertise and information to contribute to the awarding process, and their contribution to the discussion of the evidence should lead to a consensus about the placement of the level boundaries. If there is minor uncertainty which cannot be resolved through inspection of evidence, a majority view will be acceptable. If there is a major dispute, and it is not possible to resolve this with the assistance of other Board staff, the ultimate responsibility for a decision will rest with the SAC members.

2. PREPARATORY WORK

2.1 The paragraphs which follow describe the preparatory work which will be undertaken by the participants. It is presented in terms of a core which consists wholly or mainly of written components; some modification to these procedures may be needed in some other circumstances.

In the document The Award of Core Levels, and the Verification of A Level grades for the Wessex Project there is a description of materials to be circulated in advance of, and to be available at the core award (see paras. 3.2, 3.3 and 3.5). The paragraphs which follow make explicit the actions to be taken by the subject officer, the examiner and the SAC members and Wessex representative.

2.2 In advance of the meeting the subject officer will ensure that each awarder has the general package of materials (a brief description of the Wessex scheme, the syllabus and the necessary claim forms), and the core award package (this procedure paper, copies of examination papers and mark schemes and copies of materials relating to other core components).

2.3 The subject officer should collect together, for the meeting, the necessary statistical information about the core assessment components. This should include the component raw mark allocations and any scaling factors, together with component raw mark distributions (complete with means and standard deviations) and a core mark distribution. In subsequent years a summary of this type of information for the past year will also be compiled, and the subject officer will also ensure that copies of any reports or studies of relevant Wessex awards are to hand at the meeting.

2.4 The subject officer will arrange for all scripts to be on hand, arranged by centre within core component, and for there to be a suitable supply of forms for recording interim and final decisions. At later awards only a sample of scripts may be supplied.

2.5 The examiner should be prepared to comment upon the standard of the work in the core examination and to suggest some initial ideas about positions for level boundaries for each of the components (see the following sections for a discussions of the basis for these boundaries). Because of the early stage of the development of this scheme the examiner should be prepared to discuss in some detail any problems which need to be borne in mind by the awarders.

2.6 The SAC members and the Wessex representative must come to the meeting having familiarised themselves with the background materials, syllabus, question papers and marking schemes. Because of their different backgrounds, and because this is a new scheme they may wish to have an opportunity to clarify certain points at the beginning of the meeting, and the chairman should be prepared to allow a short time for this, although must confine the discussion to issues which affect the immediate issue of awarding core levels.

3. THE MEETING

3.1 In general there will be two kinds of activity at the meeting, (i) open discussion of matters which affect the award, of preliminary notions which the awarders bring to the meeting and of the conclusions reached as a result of evidence considered during the meeting, and (ii) the quiet review of evidence, primarily from scripts, conducted by each individual separately. There is no fixed agenda for the meeting, but the award will be conducted component by component, and should begin with a session which incorporates the following items

3.1.1 The Chairman should open the meeting by reminding members of the task to be undertaken, and of the developmental as well as operational nature of the Wessex Project.

3.1.2 The subject officer should ensure that the awarders are given details of mark allocations, scalings etc. (but not, at this stage the distributions of marks obtained on core components).

The Chairman will wish, at this stage to deal with general points of information, especially for the benefit of those who have had limited previous contact with the Wessex Project.

3.2 The meeting will then proceed to deal with each component in turn. The order is relatively unimportant, except that components where the evidence is not directly or wholly available to the awarders should be left to the end. The aim is to deal completely with one component before turning to another.

3.3 The examiner responsible for the component will report on candidates' responses to it, and indicate where the awarders might seek to look for level boundaries in the first instance. (The basis for these recommendations is discussed in the next section.) Awarders will review scripts in each of these regions, the subject officer managing the supply of scripts as requested.

3.4 The participants will deal with one boundary at a time. Once a reasonable time has been allowed for reading the Chairman will seek consensus about a provisional component boundary. By discussion, and perhaps by a further inspection of scripts the awarders should reach agreement, or be prepared to accept a provisional boundary which is a compromise between their various perceptions. The Chairman must not allow the meeting to stagnate over these decisions; at this stage they are provisional, and there is a further opportunity for review at the end of the meeting.

3.5 This procedure is repeated for each of the other specified boundaries for the first component, and the subject officer should maintain a record of each decision reached.

3.6 The meeting will then consider each other component. Where no material evidence is available, or where this is incomplete, the examiner may be able to describe the work done, or other members of the meeting may contribute information. Where this is a practical component awarders will make decisions based upon the criteria used for awarding marks.

3.7 The component level boundaries are added together (after scaling as required) in order to obtain boundaries for the core as a whole. The awarders must then scrutinise the total work of a number of candidates who have achieved minimum overall marks for each of the specified levels in order to satisfy themselves that the overall standard is correct. At this stage the subject officer should provide the awarders with all the statistical evidence about the components, although it is not possible for the time being to draw upon evidence from other awards. It is most important to remember that this is the award of a Wessex core and not the award of a whole A Level.

3.8 As a result of this overview the awarders may wish to modify their original decisions so that the final overall boundaries are to their satisfaction. Where some boundaries are interpolated (see below) the subject officer should complete this task so that the awarders can see the totality of the award before they disperse. The Chairman should ensure that these review procedures are conducted thoroughly and that the awarders are satisfied with the results of their deliberations. The Chairman should sign the form upon which the decisions are recorded, and the subject officer should pass this to the administrative officer responsible for the Wessex Project, retaining a copy for future reference.

4. THE BASIS FOR DECISIONS ON LEVEL BOUNDARIES

4.1 It is the intention that the award of levels with the Wessex Project shall be based entirely upon criteria in the form of attainment targets. Indeed, in some subjects much of the assessment may be conducted directly in terms of levels but where marks are used the criteria will enable awarders to fix appropriate boundaries. This basis is now established for modules, and has been incorporated into later syllabus developments, but has not yet been undertaken for the Wessex cores. It is intended that these criteria will be completed in due course, and that all Wessex subjects will be awarded with them.

4.2 The link between core and module levels and A Level grades is provided by the combination rules which are shown in the general introduction to the Wessex scheme.

4.3 The combination rules (which are provisional) are designed to reflect the 60-40 core-modules weighting and to cope with all mixes of performance between core and modules. It is obvious from the table that the core level exercises a strong influence upon the final grade so, for candidates with balanced performances across core and modules, core level 1 will lead to grade A, level 2 to B and so on. This is not invariably the case but it will serve as a basis for this first core award. Awarders are asked, therefore, to make decisions in terms of A Level grades for the time being.

4.4 Normal A Level grading concentrates on judgemental decisions at three boundaries: these are the A/B, B/C and E/N boundaries. The corresponding level boundaries are 1/2, 2/3 and 5/6. Boundaries between these are obtained by interpolation, using rules which subject officers already use for all A Level grading. These rules will apply in exactly the same way in this Wessex award. The awarders are asked to begin with the 5/6 boundary (fixed with respect to grades E and N), then to deal with the 2/3 boundary (in terms of B/C), finally considering 1/2 (A/B). The first two of these judgemental points are in

accord with the requirements of DES for all A Levels; the last is in accordance with the Board's previous practice, where this boundary was also fixed with reference to awarders' judgements of quality of work. Here all three boundaries should be treated in the same way.

4.5 The verification meeting (which follows the core award) provides an opportunity for correcting any errors which arise from this temporary procedure, or from other parts of the operation of the Wessex assessment system.

June 1989

(Revised May 1990)

wessawco2

APPENDIX 3

VERIFICATION PROCEDURES FOR THE WESSEX PROJECT

1. INTRODUCTION

1.1 The verification procedure for each Wessex Project subject is designed to provide evidence concerning the acceptability of the grades which have been awarded as a result of the combination of results from core and modules. It is a procedure which involves making judgements concerning the comparability of grading standards between the subject in the Wessex scheme and the same subject (or the nearest available equivalent) operated conventionally by the Board. For this reason the personnel conducting the verification process must

(a) be of sufficient experience, and have subject competence which would enable them to make such judgements, and

(b) be immediately familiar with the grading decisions made in the conventional subjects to which the standard of the Wessex grading is being referred.

1.2 Verification is primarily a device to ensure that the procedures for determining module and core results, and then for combining them, are producing appropriate A Level grade outcomes. If, as a result of the verification process, it is agreed that the grades generated by the Wessex procedures are comparable to those being achieved in the equivalent conventional examination(s) then those results will be published. Unless there are other reasons for making changes the assessment procedures will remain unaltered. If repeated verification procedures demonstrate that the grade outcomes are appropriate then it will be appropriate to operate verification as an occasional monitoring procedure rather than as a regular component of the Wessex scheme.

1.3 Where the Wessex grades are seen to be inappropriate there are two consequences

(a) In the longer term some changes may be needed in either the procedures or standards used to award modules or core, or in the rules for combining core and module levels. The latter is a more drastic change since it affects all Wessex subjects, and so would need to be indicated as a result of verification meetings in several subjects. Changes to the standards to be applied in making core or module awards are likely to be easier to effect.

(b) In the short term the results obtained by the candidates in the current cohort will require adjustment. Mechanisms for achieving this are given later in this paper.

1.4 Verification meetings are devices for monitoring the outcomes of the Wessex scheme, with a view to improving and refining aspects of the process of assessment. They are not awarding meetings, nor are they borderline reviews, and it is important that those participating in them do not treat them as such. For this reason it is essential that the structure of the Wessex scheme and the contents of this document are thoroughly familiar to those involved in the verification meeting.

2. PERSONNEL INVOLVED IN VERIFICATION

2.1 The verification meeting will be attended by

- two or three members of the AEB Standing Advisory Committee (SAC) for the subject, at least two of whom have been at the award of the Board's examination(s) in the subject;
- the subject co-ordinator from the Wessex Project, or an alternate who has the required subject expertise;
- the chief examiner and moderator for the subject for the Wessex Project.

2.2 The meeting will be serviced by the AEB subject officer, and other AEB staff may attend in order to advise on specific matters, or to observe the proceedings. The meeting will be chaired by one of the SAC members, appointed by the SAC to this role, or invited by the subject officer.

2.3 All of those involved in carrying through the verification process will have been supplied with a general package of materials and with a copy of this procedure paper.

2.4 Normally the verification meeting will follow the core award and the personnel will be similar for both. Both should follow closely the award for the Board subject, so that at least two of the people concerned in the verification process are directly relating Wessex achievement to that which they have just seen.

3. THE VERIFICATION PROCEDURE

3.1 The meeting will have available to it the complete work done by some or all of the Wessex candidates for the subject. This work will be assembled in sets, one per candidate, with a cover sheet which identifies the candidate by number, and shows the levels awarded for each module and for the core, with the resulting A Level grade, calculated using the Wessex combination rules.

3.2 Normally speaking the subject officer will attempt to ensure that there is work from at least three Wessex candidates in each of grades A-N. In practice it is unlikely that any distinction can be made between 'borderline' membership of a grade, and cases where a candidate's work places him or her in mid-grade, although the combination rules indicate a range of performances for each grade. A shortage of candidates in any grade may be compensated for by additional cases in adjacent grades. Where there is any choice of cases to be selected balanced performance amongst the modules is to be preferred, and

the widest possible representation of modules provided. Cases where any request for special consideration has been made or where a candidate has re-taken one or more modules and is being awarded a result for a second time, should not be included.

3.3 Also available will be some samples of scripts from the Board's conventional examination, also assembled by candidate, and representing a range of grades. Normally speaking three cases should be available for each of the grades A-N, each with reasonably balanced performances across the components and, wherever possible, drawn from the mid-point of the grade mark range or, in the case of grade A, one third of the way up from the A/B boundary towards the highest achieved overall mark. Where components such as practicals cannot be represented at the meeting a note of the performance of the candidate on this component will be included in the bundle of work.

3.4 Copies of the Wessex question papers, mark schemes and any other relevant materials will already have been sent to participants, as well as the question papers and mark schemes from the Board's conventional examination. Participants will also have a supply of forms upon which they can record their own decisions, and the subject officer will have a supply of forms upon which overall decisions and recommendations can be recorded.

3.5 It is important for the chairman to recognise that the participants will not all have the same background of knowledge and experience of the Wessex Project or of the Board's examination, and each person will need to be allowed time to 'catch up' in one respect or another. Although they will need time to confer they also need to be able to use a lot of time for reading scripts and other material. The chairman will need to ensure a proper balance of these activities so that well-informed decisions can be reached, with all of the participants contributing to these whilst not allowing the meeting to run for an undue length of time. Thus, for example, there will be some at the meeting who will not have read any Wessex materials, but who are already very familiar with the Board examinations: they will need to concentrate on Wessex candidates, making occasional references to other

materials. On the other hand, the Wessex representative and the moderator will probably be very familiar with the Wessex module materials, but will need time to look at scripts from both the Wessex core examination and at materials from the Board examination while the examiner studies some of the Wessex module materials.

3.6 The participants should take all the work which they see at face value. They are not entitled to make allowance for any special circumstances which might have affected the work of individuals: had this been a matter of significance the centre would have made a case for special consideration. Nor are they entitled to treat module materials differently because modules were taken at different times in the course. They are required to identify the features of the work which has been presented as they relate to the A Level grade awarded to the candidate.

3.7 However, they are likely to need assistance in dealing with the module work, and particularly where the volume of work is large. They cannot expect to read it all thoroughly, nor must they attempt to re-mark it. It has all been moderated and so it must be assumed that the marking has been found to be satisfactory in all respects. Part of the task of the chairman is to see that relevant information and assistance are given when needed, so that the Wessex representative and the moderator may need to inform the other participants about some aspects of the operation of the modules.

3.8 Each participant will not read all of the Wessex material. Each individual should aim to inspect at least three cases in each of grades A, B and C, and three in each of grades E and N. Even at that coverage the task is fairly daunting, and the participants are not likely to be able to read everything thoroughly, but only to look at enough to be able to gain clear impressions of the quality of the work. Unless circumstances dictate otherwise the chairman should manage the meeting so that the participants work together in one region of the grade scale, and review their decisions at intervals. It is probably best to begin by looking at candidates who have been awarded grades E and N, with a view to establishing

an idea of whether the boundary between these grades is at an appropriate standard. Having looked at this area attention should be focussed upon grades around the B/C boundary, and grade A can be looked at last of all. In the process the participants may wish to look at extra cases, to include some candidates awarded grade D, or who were ungraded, or attempt some matching of responses by Wessex candidates to those taking the Board's conventional examination. Some flexibility of approach is therefore required, but it is important that the chairman splits up the work so that some agreement is reached about one portion of the grade scale, before moving to another.

3.9 Most importantly the available time should be apportioned sensibly. The meeting must not become so immersed in one portion of the grade scale that another is left too little time; the participants are unlikely to be able to arrive at sensible decisions very late in the day. It is therefore important to recognise the value of considering one area of the grade scale, reviewing it and arriving at a provisional or tentative opinion about Wessex scheme standards at that point, then moving on to another area and treating it similarly. An overview at the end is likely to be easier as a result. At each stage, however, the chairman should see to it that each participant makes an adequate note of his or her views about the work of individual candidates, and that there is a sufficiently complete record of the provisional corporate decisions.

3.10 At the end of the meeting the whole outcome must be reviewed, and recommendations made. Here it is important to refer back to paras 1.2 and 1.3 of this procedure paper. It is for the verification meeting to recommend the action which should be taken, and a number of examples of situations which may arise will make clear what sorts of recommendations there might be.

3.10.1 The participants might agree (with only occasional exceptions) that the grading was appropriate for all cases which they have looked at. There is no reason to suppose that these cases are untypical, nor that it is unsafe to recommend that

(i) the grades of all candidates remain as calculated, and

(ii) the same procedures be used in subsequent years.

3.10.2 The participants may be content to confirm the grade, but might nevertheless feel that there is room for improvement in the procedures for determining the results. The chairman should see that recommendations are adequately recorded.

3.10.3 There may be a clear trend towards leniency or severity right through the grade range. In this case it would be important to pin-point the cause. Perhaps the awarding of core or module levels was inappropriate, or the combination rules do not satisfactorily combine these levels. It is assumed that it would be safe to alter the grades of all candidates on the basis of the sampling done during the meeting, but this would need to be accompanied by specific recommendations for changes to the scheme (but not altering its fundamental structure, which is subject to separate evaluation[1]) and sufficient additional inspection of the work of candidates to assure the participants that the trend was uniform, and applying to all.

3.10.4 It is likely that, within the three conditions described in 3.10.1 - 3.10.3 above, there will be isolated cases which do not follow the trend. Thus, although for almost all cases looked at, the participants agree that grading is correct (as in 3.10.1), one or two individuals might be perceived to be severely or leniently graded by one grade. Had this been by two grades, or had there been more than a couple of such cases, this would have cast doubt on the general perception of correct grading. In these circumstances the only course would be to inspect more cases in

[1] Thus, the use of levels and combination rules, and the core/module structure are not open to change at this stage, but the basis for level awards, and the rules of combination may be changed, although changes to the latter would only follow consistent evidence from a number of verification meetings, in various subjects, since all would be affected.

order to see whether the general perception is, in fact, sustained, whether another simple trend can be perceived, or whether the grading is actually rather erratic. The last of these possibilities will cause the greatest difficulty since it will be necessary to inspect a wide range of candidates. Several important points must be borne in mind.

(a) Assessment processes are not wholly reliable, and all examinations throw up cases which are thought to be marginal. The magnitudes of the discrepancies between awarded and perceived grades must be large enough and must happen often enough to lead to genuine doubt about the Wessex procedure.

(b) The participants in the verification meeting are not re-assessing, but are reading through responses, often rapidly, seeking to gain an overall impression. This is a good way of gaining such impressions, spread across many candidates' work, but a less good way of making certain judgements about individual grades.

(c) Concerns of this type must be widely shared amongst the participants at the meeting; it is for the chairman to decide whether the concern of one person, which is not supported by others, gives sufficient grounds for pursuing the issue.

Only, therefore, where there is certain evidence of inconsistency of grading would the meeting look to adjust grades on the basis of the inspection of all candidates, treated as individuals, and this would almost certainly require that the meeting is prolonged beyond its projected half-day session. Officers would make special arrangements to deal with this situation.

3.11 The last circumstance described above is unlikely to occur, and difficult to prescribe for in advance. Officers are expected to advise the chairman on the best courses of action, and it is most important to avoid a situation where a casual re-grading based on relatively quick impressions of work, replaces the more precise and controlled processes which have been used for core and module assessments. The pilot status of the Wessex Project must be borne in mind, with the need to interpret the outcomes of this verification procedure alongside those from other such meetings to be held in 1990 and thereafter.

3.12 The combination rules include a section which covers the grade outcomes when one or two module results are unclassified (when there is either an unclassified result on the core or more than two unclassified module results the A Level outcome is, itself, unclassified). This aspect of the combination rules may merit special attention, since it deals, in effect, with the situation where up to 20% of the assessment requirements have not been met. There is also, by implication, a very unbalanced performance by any candidate who is graded in this way. The meeting should consider any cases which fall into this category, as requiring particular consideration. It is possible that the combination rules work perfectly well in such situations, but also possible that they will require some special alteration in order adequately to deal with such unbalanced performances. It is anticipated that, in the Project pilot stages, there will be very few cases of this kind.

4. RECORDING VERIFICATION OUTCOMES

4.1 It has already been noted that the chairman is responsible for seeing that an adequate record of the decisions is made; in practice the subject officer will probably undertake this task, and must ensure that the records of decisions made are collected at the end of the meeting. In broad terms these will relate to

(a) decisions about grade changes to the current cohort of candidates, and

(b) recommendations for changes to the procedures for obtaining grades.

4.2 The first of these outcomes will be implemented by the Board officer responsible for the administration of the Wessex Project assessment. The second will be held by the Board's Development Officer for the Project, who will consider the changes which are to be made in consultation with other officers and with Project staff. The implementation of these changes will follow at the earliest possible date following the meeting, but this is unlikely to be until the Autumn.

4.3 A review of the outcomes of the meeting will also form part of the evidence to be incorporated into the evaluation of the examinations for the Project, to be assembled by the Board's Research Officer who is responsible for this aspect of the work. In due course, this information will form part of reports to SEAC, organised according to the agreed evaluation procedures.

June 1989

wessex

APPENDIX 4

FORMS TO BE USED IN CONNECTION WITH THE VERIFICATION OF A LEVEL GRADES FOR THE WESSEX PROJECT

1. Cover Sheet to be attached to the complete work of each candidate.

2. Form to be used by participants for recording their opinions concerning the grades appropriate for individual candidates.

3. Form for recording the decisions of all participants concerning each candidate: for chairman's or servicing officer's use.

FORM TO BE USED TO RECORD CORE AWARD DECISIONS

4. Form for recording final decisions (also to be used to record interim component decisions).

THE ASSOCIATED EXAMINING BOARD

THE WESSEX PROJECT VERIFICATION MEETINGS

FORM 1: COVER SHEET

Subject:

Examination Session:

Centre No.

Candidate No.

Candidate Name:

	Number	Level Awarded
First Module Second Module Third Module Fourth Module		
Core		

A Level grade from combination rules

THE ASSOCIATED EXAMINING BOARD

THE WESSEX PROJECT VERIFICATION MEETINGS

FORM 2: PARTICIPANTS' RECORD

SUBJECT _____ EXAMINATION SESSION _____

Centre No.	Cand. No.	Grade Awarded*	Participants' grade/notes

*from combination rules, and appearing on cover sheet to each candidate's work.

THE ASSOCIATED EXAMINING BOARD

THE WESSEX PROJECT VERIFICATION MEETINGS

FORM 3: MEETING DECISIONS

SUBJECT _____ EXAMINATION SESSION _____

Centre No.	Cand. No.	Grade Awarded*	Participants' Grades							Agreed Grade
			1	2	3	4	5	6	7	

Chairman's signature_____

* from combination rules

THE ASSOCIATED EXAMINING BOARD

THE WESSEX PROJECT CORE AWARD

FORM 4: AWARDING DECISIONS

Lowest acceptable mark for Level	Component Levels			Overall Core Levels
	1	2	3	
1				
2				
3				
4				
5				
6				

Chairman's Signature_____

wesawd

APPENDIX 7.2

AGGREGATION RULE STUDY DOCUMENTS

Judge's Name .....

FORM I

Research into Wessex Chemistry Combination Rules

GRADE A

Centre	Candidate	Clearly Grade A	Just Grade A	Not quite Grade A	Clearly not Grade A

GRADE B

Centre	Candidate	Clearly Grade B	Just Grade B	Not quite Grade B	Clearly not Grade B

GRADE E

Centre	Candidate	Clearly Grade E	Just Grade E	Not quite Grade E	Clearly not Grade E

Judge's Names .....

.....

FORM C

Research into Wessex Chemistry Combination Rules

GRADE A

Centre	Candidate	Agreement not reached	Clearly Grade A	Just Grade A	Not quite Grade A	Clearly not Grade A

GRADE B

Centre	Candidate	Agreement not reached	Clearly Grade B	Just Grade B	Not quite Grade B	Clearly not Grade B

GRADE E

Centre	Candidate	Agreement not reached	Clearly Grade E	Just Grade E	Not quite Grade E	Clearly not Grade E

APPENDIX 8.1

EXPERIMENTAL EXAMINATION ATTAINMENT TARGET

SUMMARY STATISTICS AND LEVEL BOUNDARY MARKS

Tier 4 - 6 (86 candidates)		Paper 1	Paper 2	Coursework	Overall
Ma1	max. mark			20	20
	mean			4.8	4.8
	SD			2.5	2.5
Ma2	max. mark	14	16		30
	mean	5.2	5.8		11.0
	SD	2.7	2.8		4.9
Ma3	max. mark	15	14		29
	mean	6.7	3.3		10.0
	SD	3.0	2.1		4.3
Ma4	max. mark	15	16		31
	mean	2.3	4.8		7.2
	SD	2.4	2.4		3.9
Ma5	max. mark	16	14		30
	mean	6.4	2.8		9.2
	SD	3.5	2.6		5.4
Total	max. mark	60	60	20	150
(Ma1 scaled x 1.5)	mean	20.6	16.8	4.8	44.8
	SD	8.5	7.4	2.5	16.8

Tier 4 - 6 Level boundaries (lowest mark in each level)

Level	Ma1	Ma2	Ma3	Ma4	Ma5
4	1	8	7	7	8
5	4	16	14	12	15
6	7	24	21	17	22

Tier 5 - 8 (275 candidates)		Paper 1	Paper 2	Coursework	Overall
Ma1	max. mark			20	20
	mean			9.7	9.7
	SD			3.7	3.7
Ma2	max. mark	22	22		44
	mean	10.8	12.0		22.8
	SD	4.6	4.9		8.7
Ma3	max. mark	23	22		45
	mean	10.9	8.6		19.6
	SD	5.3	3.2		7.8
Ma4	max. mark	22	23		45
	mean	8.8	8.0		16.8
	SD	5.8	4.8		9.8
Ma5	max. mark	23	23		46
	mean	8.7	7.6		16.3
	SD	5.0	3.9		8.1
Total (Ma1 scaled x 2.25)	max. mark	90	90	20	225
	mean	39.2	36.3	9.7	97.5
	SD	17.3	13.6	3.7	35.9

Tier 5 - 8 Level boundaries (lowest mark in each level)

Level	Ma1	Ma2	Ma3	Ma4	Ma5
5	4	4	6	5	6
6	7	18	18	16	16
7	12	31	29	27	26
8	16	36	38	35	33

Tier 7 - 10 (128 candidates)		Paper 1	Paper 2	Coursework	Overall
Ma1	max. mark			20	20
	mean			15.0	15.0
	SD			3.2	3.2
Ma2	max. mark	24	20		44
	mean	11.3	9.2		30.6
	SD	4.3	3.8		6.9
Ma3	max. mark	21	24		45
	mean	8.7	7.5		16.3
	SD	4.7	4.3		8.3
Ma4	max. mark	24	24		48
	mean	12.3	10.4		22.7
	SD	6.4	6.2		11.8
Ma5	max. mark	21	22		43
	mean	11.3	6.3		17.5
	SD	4.4	4.6		7.9
Total	max. mark	90	90	20	225
(Ma1 scaled x 2.25)	mean	43.7	33.4	15.0	110.7
	SD	16.7	14.9	3.2	35.9

Tier 7 - 10 Level boundaries (lowest mark in each level)

Level	Ma1	Ma2	Ma3	Ma4	Ma5
7	12	5	3	4	7
8	16	16	14	16	16
9	19	28	25	29	25
10	20	39	33	40	38

APPENDIX 8.2

CROSS-TABULATIONS (NUMBERS OF CANDIDATES) OF WHOLE SUBJECT LEVELS FROM THE TWO AWARDING PROCEDURES WITH TEACHERS' ESTIMATES.

Tier 4 - 6

		Strongly criterion-referenced level				
		U	4	5	6	Total
Estimated level	U					0
	4	50	1	1		52
	5	20	3	0		23
	6	2	8	1		11
Total		72	12	2	0	86

Tier 4 - 6

		Conventional level				
		U	4	5	6	Total
Estimated level	U					0
	4	16	34	2		52
	5	3	7	13		23
	6		2	8	1	11
Total		19	43	23	1	86

Tier 5 - 8

		Strongly criterion-referenced level					
		U	5	6	7	8	Total
Estimated level	U	8					8
	5	29	1				30
	6	87	5	3	4		99
	7	32	18	36	40	1	127
	8		1	2	3	5	11
Total		156	25	41	47	6	275

Tier 5 - 8

		Conventional level					
		U	5	6	7	8	Total
Estimated level	U	3	5				8
	5	1	21	8			30
	6		53	41	5		99
	7		3	81	40	3	127
	8			1	2	8	11
Total		4	82	131	47	11	275

Tier 7 - 10

		Strongly criterion-referenced level					
		U	7	8	9	10	Total
Estimated level	U	6					6
	7	37	6	3			46
	8	27	11	8	3		49
	9	3	2	8	2		15
	10			3	9		12
Total		73	19	22	14	0	128

Tier 7 - 10

		Conventional level					
		U	7	8	9	10	Total
Estimated level	U	1	3	2			6
	7	1	39	5	1		46
	8		5	40	4		49
	9		1	3	10	1	15
	10				9	3	12
Total		2	48	50	24	4	128

REFERENCES

Adams, R M (1993) *Personal Communication*

Adams, R M and Murphy, R J L (1982) The achieved weight of examination components. *Educational Studies* 8(1) 15 - 22

Adams, R M and Wilmut, J (1981) A measure of the weights of examination components and scaling to adjust them. *The Statistician* 30 263 - 269

AEB (1995) *Statistics Summer 1994* (Guildford; Associated Examining Board)

Aldrich, V C (1963) *Philosophy of Art* (Englewood Cliffs; Prentice-Hall)

Apple, M W (1978) Ideology and Educational Reform. *Comparative Education Review* 26 367-387

Ayer, A J (1946) *Language Truth and Logic* Second edition (Harmondsworth; Penguin)

Back, K W (1951) Influence through social communication. *Journal of Abnormal and Social Psychology* 46 9 - 23

Bakeman, R & Gottman, J M (1986) *Observing Interaction: An Introduction to Sequential Analysis.* (Cambridge; Cambridge University Press)

Bales, R F (1950) *Interaction Process Analysis.* (Reading; Mass.: Addison-Wesley)

Bales, R F (1970) *Personality and Interpersonal Behaviour.* (New York; Holt, Rinehart & Winston)

Bambrough, J R (1979) *Moral Scepticism and Moral Knowledge* (London; Routledge)

Bardell, G S; Forrest, G M and Shoesmith, D J (1978) *Comparability in GCE: a review of the boards' studies, 1964 - 1977* (Manchester; Joint Matriculation Board)

Bardell, G; Fearnley, A and Fowles, D (1984) *The contribution of graded objectives schemes in Mathematics and French* (Manchester; Joint Matriculation Board)

Baron, R S; Kerr, N & Miller, N (1992) *Group Process, Group Decision, Group Action* (Buckingham; Open University Press)

Beardsley, M C (1981) *Aesthetics: Problems in the Philosophy of Criticism* (Indianapolis; Hackett)

Becker, H S (1970) Problems in Inference and Proof in Participant Observation - in Filstead, W J (1970) *Qualitative Methodology* (Chicago; Markham)

Best, D (1985) *Feeling and Reason in the Arts* (London; Allen & Unwin)

Billington, R (1988) *Living Philosophy: An Introduction to Moral Thought* (London; Routledge)

Bishop, K; Bullock, K; Martin, S and Thompson, J (1996) *Users Perceptions of the GCSE* (Manchester; The Joint Council for the GCSE)

Blackstone, T and Mortimore, J (1982) *Disadvantage and Education.* (London; Heinemann)

Bogdan, R C & Biklen, S K (1982) *Qualitative Research for Education* (Boston; Allyn and Bacon)

Braun, H.I. and Holland, P.W. (1982) Observed-score test equating: a mathematical analysis of some ETS equating procedures. In Holland, P.W. and Rubin, D.B. (1982) *Test Equating* (New York; Academic Press)

Brimer, A; Madaus, G F; Chapman, B; Kellaghan, T and Wood, R (1978) *Sources of Difference in School Achievement* (Windsor; NFER Publishing Company)

Broadfoot, P (1979) *Assessment, Schools and Society* (London; Methuen)

Broadfoot, P (1986) Alternatives to Public Examinations in Nuttall, D L (Ed) (1986) *Assessing Educational Achievement* (London; Falmer)

Brown, R (1988) *Group Processes: Dynamics within and between Groups* (Oxford; Blackwell)

Choppin, B and Orr, L (1976) *Aptitude testing at eighteen plus.* (Slough; NFER)

Christie, T (1982) *The Marking and Grading of Examinations* Invited paper presented at an Associated Examining Board Research Seminar in December 1982. (Guildford; AEB)

Christie, T and Forrest, G M (1981) *Defining Public Examination Standards.* (London; Schools Council/Macmillan)

Cizek, G J (1993) Reconsidering Standards and Criteria. *Journal of Educational Measurement* 30(2) 93 - 106

Collingridge, D (1982) *Critical Decision Making: A New Theory of Social Choice.* (London; Pinter)

Cresswell, M J (1986a) A Review of Borderline Reviewing. *Educational Studies* 12(2) 175 - 190

Cresswell, M J (1986b) Examination Grades: how many should there be? *British Educational Research Journal* 12(1) 37 - 54

Cresswell, M.J. (1987a) A more generally useful measure of the weight of examination components. *British Journal of Mathematical and Statistical Psychology.* 40(1) 61 - 79

Cresswell, M J (1987b) Describing Examination Performance: grade criteria in public examinations *Educational Studies* 13(3) 247 - 265

Cresswell, M.J. (1987c) *Grade Criteria: Some Unresolved Issues.* Invited paper presented at the H.M.I. Conference: *GCSE Grade Criteria.* Solihull, 18 - 19 March.

Cresswell, M J (1988) Combining Grades from Different Assessments - How Reliable is the result? *Educational Review* 40(3) 361 - 383

Cresswell, M J (1990) Wessex modular A-levels in 1990: a technical evaluation *AEB Research Report RAC/513* (Guildford; Associated Examining Board)

Cresswell, M J (1993a) Pre-publication version of Cresswell (1994) presented at the School Curriculum and Assessment Authority Conference: *Issues of Reliability in Criterion-Related National Curriculum Tests* (London; 8 December 1993).

Cresswell, M J (1993b) A Mulitlevel Bivariate Model in Prosser, R; Rasbash, J and Goldstein, H (1993) *Data analysis with ML3* (London; Institute of Education)

Cresswell, M J (1994) Aggregation and Awarding methods for National Curriculum Assessments in England and Wales: a comparison of approaches proposed for Key Stages 3 and 4. *Assessment in Education* 1(1) 45 - 61

Cresswell, M J (1995) Technical and Educational Implications of using Public Examinations for Selection to Higher Education. in Kellaghan, T (Ed) (1995) *Admission to Higher Education: Issues and Practice* (Dublin; Educational Research Centre and Princeton; International Association for Educational Assessment)

Cresswell, M J (1996) Defining, Setting and Maintaining Standards in Curriculum Embedded Examinations: Judgemental and Statistical Approaches in Goldstein, H and Lewis, T (1996) *Assessment: Problems, Developments and Statistical Issues* (London; Wiley)

Cresswell, M J; Elwood, J; Macdonald, H and Patrick, H (1993) Inter-Group Research Committee Key Stage 4 Pilot Examinations: Interim Report on Aggregation and Awarding. *(Unpublished report for the Joint Council for the GCSE and the School Examinations and Assessment Council; March 1993)*

Cresswell, M J and Gubb, J (1987) *The Second International Mathematics Study in England and Wales* (Windsor; NFER-Nelson)

Cresswell, M J and Houston, J G (1991) Assessment of the National Curriculum - some fundamental considerations. *Educational Review* 43 63 - 78

Dearing, R (1994) *The National Curriculum and its Assessment: Final Report (December 1993)* (London; School Curriculum and Assessment Authority)

Dearing, R (1996) *Review of Qualifications for 16 - 19 Year Olds* (London; School Curriculum and Assessment Authority)

Delap, M (1994) The interpretation of the achieved weights of examination components. *The Statistician* 43(4) 505 - 511

Dennett, D (1993) *Consciousness Explained* (London; Penguin)

DES (1979) *Aspects of Secondary Education: A survey by HM Inspectors of Schools.* (London; HMSO)

DES (1982) *Examinations at 16-plus: a statement of policy* (London; Department of Education and Science/Welsh Office)

DES (1991) *Mathematics in the National Curriculum* (London; HMSO)

Deutsch, M & Gerard, H B (1955) A study of normative and informational social influence upon individual judgement. *Journal of Abnormal and Social Psychology* 51 629 - 636

Dore, R (1976) *The Diploma Disease* (London; Unwin Educational)

Edgeworth, F Y (1890) The elements of chance in competitive examinations Parts I and II. *Journal of the Royal Statistical Society* 53 460 - 475 and 644 - 663

Eiser, J R (1990) *Social Judgement* (Milton Keynes; Open University Press)

Everitt, B (1977) *The Analysis of Contingency Tables* (London; Chapman and Hall)

Fitz-Gibbon, C T and Vincent, L (1994) *Candidates' performance in public examinations in Mathematics and Science* (London; School Curriculum and Assessment Authority)

Fogelin, R J (1967) *Evidence and Meaning: Studies in Analytic Philosophy* (London; Routledge)

Forrest, G M (1995) Admission to Higher Education in England and Wales: Retrospect and Prospect in Kellaghan, T (Ed) (1995) *Admission to Higher Education: Issues and Practice* (Dublin; Educational Research Centre and Princeton; International Association for Educational Assessment)

Forrest, G M and Orr, L (1984) *Grade Characteristics in English and Physics* (Manchester; Joint Matriculation Board)

Forrest, G M and Shoesmith, D (1985) *A second review of comparability studies* (Manchester; Joint Matriculation Board)

Foxman, D; Ruddock, G; Joffe, L.; Mason, K; Mitchell, P and Sexton, B. (1985) *A Review of Monitoring in Mathematics 1978 to 1982* (London; Assessment of Performance Unit)

French, S; Slater, J B; Vassiloglou, M; and Willmott, A S (1987) *Descriptive and Normative Techniques in Examination Assessment* (Oxford; UODLE)

Gipps, C (1990) *Assessment: A Teachers' Guide to the Issues.* (London; Hodder and Stoughton)

Gipps, C and Murphy, P (1994) *A Fair Test? Assessment, achievement and equity* (Buckingham; Open University Press)

Glaser, B G & Strauss, A L (1967) *The Discovery of Grounded Theory* (Chicago; Aldine)

Glaser, R (1963) Instructional technology and the measurement of learning outcomes. *American Psychologist* 18 519 - 521

Glass, G V (1978) Standards and Criteria *Journal of Educational Measurement* 15(4) 237 - 261

Goacher, B (1984) *Selection Post-16: the role of examination results.* Schools Council Examinations Bulletin 45 (London; Methuen)

Goldstein, H (1983) Measuring Changes in Educational Attainment Over Time: Problems and Possibilities. *Journal of Educational Measurement* 20 69 - 378

Goldstein, H (1986a) Gender Bias and Test Norms in Educational Selection. *BERA Research Intelligence* May 2 - 4

Goldstein, H (1986b) Models for Equating Test Scores and for Studying the Comparability of Public Examinations. In Nuttall, D L (Ed) (1986) *Assessing Educational Achievement* (London; Falmer)

Goldstein, H (1991) Better Ways to Compare Schools? *Journal of Educational Statistics* 16 89 - 91

Goldstein, H (1993) Assessing Group Differences *Oxford Review of Education* 19(2) 141 - 150

Goldstein, H (1995) *Multi-level Statistical Models: Second Edition* (London; Arnold)

Goldstein, H (1996) Group Differences and Bias in Assessment in Goldstein, H and Lewis, T (1996) *Assessment: Problems, Developments and Statistical Issues* (London; Wiley)

Goldstein, H and Cresswell, M J (1996) The comparability of different subjects in public examinations: a theoretical and practical critique *Oxford Review of Education* (in press).

Goldstein, H; Rasbash, J; Yang, M; Woodhouse, G; Pan, H; Nuttall, D and Thomas, S (1993) A Multilevel Analysis of School Examination Results *Oxford Review of Education* 19(4) 425 - 433

Good, F J and Cresswell, M J (1988a) *Grading the GCSE* (London; Secondary Examinations Council)

Good, F J and Cresswell, M J (1988b) *Differentiated Assessment: Grading and Related Issues.* (London, Secondary Examinations Council)

Gray, J; Jesson, D and Jones, B (1986) The search for a fairer way of comparing schools' examination results. *Research Papers in Education* 1 91 - 122

Gray, J; Jesson, D; Goldstein, H; Hedger, K and Rasbash, J (1995) A Multi-level Analysis of School Improvement: Changes in Schools' Performance over Time *School Effectiveness and School Improvement* 6(2) 97 - 114

Guilford, J P and Fruchter, B (1973) *Fundamental Statistics in Psychology and Education; fifth edition* (New York; McGraw-Hill)

Guskey, T R and Kifer, E W (1989) *Ranking School Districts on the basis of Statewide Test Results: Is it meaningful or misleading?* Paper presented at American Educational Research Association Conference in San Francisco, March 1989

Hammersley, M (1989) *The Dilemma of Qualitative Method* (London; Routledge)

Hammond, K R; Stewart, T R; Brehmer, B and Steinmann, D O (1975) Social-Judgement Theory in Kaplan, M F and Schwartz, S (1975) *Human Judgement and Decision Processes* (New York; Academic Press)

Harrison, A (1983) *Profile Reporting of Examination Results.* (London; Methuen)

Heywood, J (1989) *Assessment in Higher Education.* (Chichester; John Wiley & Sons)

Hively, W (1970) Domain-referenced achievement testing. *Paper given at the 1970 American Educational Research Association convention.*

HMI (1992) *GCSE Examinations: Quality and Standards* (London; Department for Education)

Hofstadter, D R (1980) *Godel, Escher, Bach: an Eternal Golden Braid* (London; Penguin)

Holland, P W and Rubin, D B (1982) *Test Equating* (New York; Academic Press)

Houston, J G (1980) *Report of the Inter-board Cross-moderation study in English Literature at Ordinary Level: 1975* (Aldershot; Associated Examining Board)

Houston, J G (1987) Advanced Level Grades as Predictors of Degree Performance. *Paper prepared as part of the GCE Examining Boards' evidence to a Government Committee of Inquiry into A-level examinations chaired by Professor G Higginson during 1987/8.*

Hunter, J E and Schmidt, F L (1976) A critical analysis of the statistical and ethical implications of various definitions of test bias. *Psychological Bulletin* 83 1053 - 1071

Hutchison, D and Schagen, I (1994) *How Reliable is National Curriculum Assessment?* (Windsor; NFER)

IAASE (1993) *Report on the Work of the IAASE 1992 - 1993* (London; Independent Appeals Authority for School Examinations)

IGRC (1992) *Differentiation in GCSE Mathematics* (London; Secondary Examinations and Assessment Council)

Ingenkamp, K (1977) *Educational Assessment.* (Windsor; NFER)

Johnson, S and Cohen, L (1983) *Investigating Grade Comparability through Cross-moderation* (London; Schools Council)

Kaplan, M F and Schwartz, S (1975) *Human Judgement and Decision Processes* (New York; Academic Press)

Kellaghan, T (1995) *Admission to Higher Education: Issues and Practice* (Dublin; Educational Research Centre and Princeton; International Association for Educational Assessment)

Kahneman, D and Miller, D T (1986) Norm theory: Comparing reality to its alternatives. *Psychologogical Review* 93 136 - 153

Klein, J (1970) *Working with Groups: The Social Psychology of Discussion and Decision* (London; Hutchinson)

Lindblom C (1965) *The Intelligence of Democracy* (London; Collier-Macmillan)

Long, H.A. (1985) *Experience of the Scottish Examinations Board in developing a grade-related criteria system of awards.* Paper presented at the 11th annual conference of the International Association for Educational Assessment held in Oxford, England.

Lundy I (1993) Applying the Criterion-Related Method for Determining Cut-offs to KS3 Maths. Paper presented at the School Curriculum and Assessment Authority Conference: *Issues of Reliability in Criterion-Related National Curriculum Tests* (London; 8 December 1993).

Macdonald, H (1992) Pass Rates in Religious Studies A-level in 1990 and 1991: What Happened? *Unpublished Research Report: RAC/583* (Guildford; Associated Examining Board)

Macdonald, H (1993) IGRC Key Stage 4 Pilot Examinations 1992/1993: Final Report on Mathematics. *Unpublished report for the Joint Council for the GCSE, June 1993*

Massey, A J (1993) *Applying a "criterion-related" process for fixing total mark thresholds for NC levels to 1993 KS3 Science national test data.* Paper presented at the School Curriculum and Assessment Authority Conference: *Issues of Reliability in Criterion-Related National Curriculum Tests* (London; 8 December 1993).

Massey, A J (1995) Criterion-related Test Development and National Test Standards. *Assessment in Education* 2(2) 187 - 204

Mathews, J C (1985) *Examinations - A Commentary.* (London; George Allen and Unwin)

Messick, S (1987) *Validity.* (Princeton; Educational Testing Service)

Miles, H B (1979) *Some Factors Affecting Attainment at 18+.* (Oxford; Pergamon)

Morrison, H G; Healy, J; Busch, J C and D'Arcy, J (1993a) *Setting test standards using professional judgement: an analysis of the NISEAC Common Assessment Instrument.* Paper presented at the School Curriculum and Assessment Authority Conference: *Issues of Reliability in Criterion-Related National Curriculum Tests* (London, 8 December 1993).

Morrison, H G; Busch, J C and D'Arcy, J (1993b) *Reconciling end-of-Key Stage test scores and classroom-based assessment: a role for the Angoff procedure in British assessment.* Paper presented at the School Curriculum and Assessment Authority Conference: *Issues of Reliability in Criterion-Related National Curriculum Tests* (London, 8 December 1993).

Mortimore, J and Mortimore, P (1984) Secondary School Examinations: the helpful servants, not the dominating master. *Bedford Way Paper No 18* (London; University of London Institute of Education)

Murphy, R J L (1978) Reliability of marking in eight GCE examinations *British Journal of Educational Psychology* 48 196 - 200

Murphy, R J L (1982a) A further report of investigations into the reliability of marking GCE examinations. *British Journal of Educational Psychology* 52 58 - 63

Murphy, RJL (1982b) Sex differences in Objective Test performance. *British Journal of Educational Psychology* 52 213 - 19

Murphy, R J L;Burke, P; Cotton, T; Hancock, J; Partington, J; Robinson, C; Tolley, H; Wilmut, J and Gower, R (1996) *The Dynamics of GCSE Awarding: Report of a project conducted for the School Curriculum and Assessment Authority* (London; SCAA)

Newbould, C A and Massey, A J (1979) *Comparability using a common element* (Cambridge; University of Cambridge Local Examinations Syndicate)

Newton, P (1996) The reliability of marking of GCSE scripts: Mathematics and English *British Educational Research Journal* 22 405 - 420

Newton P (1996) Examining Standards Over Time *Research Papers in Education* (in press)

Nisbett, R E; Borgida, E; Crandall, R and Reed, H (1982) Popular induction: Information is not necessarily informative. In Kahneman, D; Slovic, P, and Tversky, A (1982) *Judgment under uncertainty: Heuristics and biases* (Cambridge; Cambridge University Press)

Nisbett, R E and Wilson, T D (1977) Telling more than we know: Verbal reports on mental processes. *Psychological Review* 84 231 - 259

Nuttall, D L (1986) Problems in the Measurement of Change in Nuttall, D L (ed) (1986) *Assessing Educational Achievement.* (London; Falmer Press)

Nuttall, D L (1987) The validity of assessments *European Journal of Psychology of Education* 11(2) 109 - 118

Nuttall, D L & Armitage, P (1984) *A feasibility Study of a Moderating Instrument.* Report to the Business and Technician Education Council

Nuttall, D L; Goldstein, H; Prosser, R and Rasbash, J (1989) Differential School Effectiveness. *International Journal of Educational Research* 13 769 - 776

Nuttall, D L and Willmott, A S (1972) *British Examinations: Techniques of Analysis.* (Windsor; National Foundation for Educational Research)

Orr, L and Forrest, G M (1984) *Investigation into the relationship between grades and assessment objectives in History and English examinations* (Manchester; Joint Matriculation Board)

Orr, L and Nuttall, D L (1983) *Determining Standards in the Proposed Single System of Examinations at 16+* (London; Schools Council)

Osburn, H G (1968) Item sampling for achievement testing *Educational and Psychological Measurement* 28 95 - 104

Pilliner, A E G (1979) Norm-referenced and Criterion-referenced Tests - An Evaluation *Issues in Educational Assessment* (Edinburgh; Scottish Education Department/HMSO)

Pirsig, R M (1974) *Zen and the Art of Motorcycle Maintenance: An Inquiry into Values* (London; Bodley Head)

Pirsig, R M (1991) *Lila: An Inquiry into Morals* (London; Bantam)

Pole, D (1961) *Conditons of Rational Inquiry: A Study in the Philosophy of Value* (London; Athlone)

Pollitt, A (1993) *Setting standards in KS3 English.* Paper presented at the School Curriculum and Assessment Authority Conference: *Issues of Reliability in Criterion-Related National Curriculum Tests* (London; 8 December 1993).

Pollitt, A; Entwistle, N; Hutchinson, C and DeLuca, C (1985) *What Makes Exam Questions Difficult?* (Edinburgh; Scottish Academic Press)

Quadling, D (1992) Making the Grade *The Mathematical Gazette* Vol.76, No.476 261 - 268

Quinlan, M (1993) *Delta Index* Paper given at Inter-Group Research Committee seminar on *Interpreting Examination Statistics* held at the offices of the University of London Examinations and Assessment Council; March.

Rachlin, H (1989) *Judgement, Decision and Choice: A cognitive/behavioural synthesis* (New York; Freeman)

Ruddock, G J and Tomlins, B with Mason, K; Holding, B; Reiss, M; Keys, W; Foxman, D and Schagen, I (1993) *Evaluation of National Curriculum Assessments at Key Stage 3: report on the 1992 national pilot assessment of mathematics and science* (Unpublished research report; School Examinations and Assessment Council)

Sadler, D R (1985) The origins and functions of evaluative criteria. *Educational Theory* 35 285 - 297

Sadler, D R (1987) Specifying and promulgating achievement standards. *Oxford Review of Education* 13 191 - 209

Sadler, D R (1989) Formative assessment and the design of instructional systems. *Instructional Science* 18 119 - 144

Savile, A (1972) The place of intention in the concept of art. in Osborne, H (Ed) (1972) *Aesthetics* (Oxford; Oxford University Press)

SCAA (1994) GCE *A and AS Code of Practice* (London; School Curriculum and Assessment Authority)

SCAA (1995) *GCSE Mandatory Code of Practice* (London; School Curriculum and Assessment Authority)

Schools Council (1979) *Standards in Public Examinations: Problems and Possibilities* Report from the Schools Council Forum on Comparability (London; Schools Council)

Sears, P S (1940) Levels of Aspiration in Academically Successful and Unsuccessful Children. *Journal of Abnormal and Social Psychology* 35 498 - 536

SEC (1984) *The development of Grade-related Criteria for the General Certificate of Secondary Education - a briefing paper for working parties.* (London; Secondary Examinations Council)

SEC (1985a) *The Development of Grade Criteria for the General Certificate of Secondary Education* (London; Secondary Examinations Council)

SEC (1985b) *Reports of the Grade-related Criteria Working Parties* (London; Secondary Examinations Council)

SEC (1986) Draft Grade Criteria *SEC News Number 2* (London; Secondary Examinations Council)

SEC (1987) Grade Criteria - Progress Report *SEC News Number 6* (London; Secondary Examinations Council)

SEG (1988) *Guidance paper for setters and revisers* (Unpublished; Southern Examining Group)

Shaw, D G; Huffman, M D and Haviland, M G (1987) Grouping Continuous Data in Discrete Intervals: Information Loss and Recovery. *Journal of Educational Measurement* 24(2) 167 - 173

Siegel, S and Castellan, N J (1988) *Nonparametric Statistics for the Behavioural Sciences: Second Edition* (New York; McGraw-Hill)

SRAC (1990) *Identification of Good Practice in GCE Examinations* (Unpublished; Standing Research Advisory Committee of the GCE Boards)

Stobart, G; Elwood, J and Quinlan, M (1992) Gender Bias in Examinations *British Educational Research Journat* 18(3) 261 - 276

Stolnitz, J (1973) The Artistic Values in Aesthetic Experience. *Journal of Aesthetics and Art Criticism* 32

Tattersall, K (1994) The Role and Functions of Public Examinations. *Assessment in Education* 1(3) 293 - 304

TGAT (1988) *Task Group on Assessment and Testing: A Report* (London; Department of Education and Science)

Thyne, J M (1974) *Principles of Examining.* (London; ULP)

Tymms, P B and Fitz-Gibbon, C T (1990) A comparison of exam boards: 'A' levels. *Oxford Review of Education* 17 17 - 32

Tversky A and Kahneman D (1974) Judgement under uncertainty: Heuristics and biases *Science* 185 1124 - 1131

Tymms, P B and Vincent, L (1995) *Comparing Examination Boards and Syllabuses at A-level: students' grades, attitudes and perceptions of classroom processes: Technical Report* (Belfast; Northern Ireland Council for the Curriculum, Examinations and Assessment)

Upshaw, H S (1975) Judgement and Decision Processes in the formation and change of social attitudes in Kaplan, M F and Schwartz, S (1975) *Human Judgement and Decision Processes* (New York; Academic Press)

Ward, C (1980) D*esigning a scheme of Assessment* (Cheltenham; Stanley Thornes)

Wesman, A G (1971) Writing the Test Item. In Thorndike, R L (1971) *Educational Measurement* (Washington DC; American Council on Education)

White, P A (1988) Knowing more about what we can tell: 'Introspective access' and causal report accuracy 10 years later. *British Journal of Psychology* 79 13 - 45

Wiliam, D (1993) *Setting cut-scores in national curriculum assessment.* Paper presented at the School Curriculum and Assessment Authority Conference: *Issues of Reliability in Criterion-Related National Curriculum Tests* (London; 8 December 1993).

Wiliam, D (1995a) Combination, Aggregation and Reconciliation: evidential and consequential bases. *Assessment in Education* 2(1) 53 - 74

Wiliam, D. (1995b) Technical Issues in Criterion-Referenced Assessment: Evidential and Consequential Bases. in Kellaghan, T (Ed) (1995) *Admission to Higher Education: Issues and Practice* (Dublin; Educational Research Centre and Princeton; International Association for Educational Assessment)

Willmott, A S. (1980) *Twelve years of Examinations Research: ETRU, 1965-1977* (London; Schools Council)

Willmott, A S and Nuttall, D L (1975) *The reliability of Examinations at 16+* (London; Macmillan)

Wilmut, J (1981) *A brief report on two factors which affect grade changes in mark-remark and weighting exercises* Associated Examining Board Research Report RAC/184 (Guildford; AEB)

Wilmut, J (1986) *Marking Reliability: A Bibliography.* Associated Examining Board Research Report RAC/403 (Guildford; AEB)

Wilmut, J (1990) *Personal Communication*

Wilmut, J and Rose, J (1989) *The Modular TVEI Scheme in Somerset: its concept, delivery and administration* Report to the Training Agency of the Department of Employment, London

Wine, J D (1982) Evaluation Anxiety: A cognitive-attentional construct in Krohne H W and Laux, L (1982) *Achievement, Stress and Anxiety.* (Washington; Hemisphere)

Wolf, A (1993) *Assessment Issues and Problems in a Criterion-based System* (London; Further Education Unit)

Wood, R (1986) Aptitude testing is not an engine for equalising educational opportunity. *British Journal of Educational Studies* 34(1) 26 - 37

Wood, R (1991) *Assessment and Testing: A survey of research* (Cambridge; Cambridge University Press)

Woodhouse, G and Goldstein, H (1988) *Educational Performance Indicators and LEA League Tables.* (London; University of London Institute of Education)

Young, M (1961) *The Rise of the Meritocracy 1870-2033* (London; Pelican)

Znaniecki, F (1934) *The Method of Sociology* (New York; Farrar & Rinehart)